Detecting *de novo* Mutations in Intellectual Disability

Joep de Ligt

DETECTING *DE NOVO* MUTATIONS IN INTELLECTUAL DISABILITY

The research presented in this thesis was performed at the Department of Human Genetics, Radboudumc, Nijmegen, The Netherlands, with financial support from TechGene EU FP7 (Health-F5-2009-223143). The printing of this thesis was kindly sponsored by; the Department of Human Genetics Radboudumc, the Radboud University and Life Technologies, which is greatly appreciated.

ISBN/EAN: 978-94-6203-549-2

Cover: adapted from efstig.deviantart.com

Printing: CPI Koninklijke Wöhrmann,The Netherlands

Detecting *de novo* Mutations in Intellectual Disability

Proefschrift
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 7 april 2014
om 14.30 uur precies

door
Joep de Ligt
geboren op 20 september 1986
te Boskoop

Promotoren:

Prof. dr. H.G. Brunner
Prof. dr. J.A. Veltman

Copromotoren:

Dr. L.E.L.M. Vissers
Dr. J.Y. Hehir-Kwa

Manuscriptcommissie:

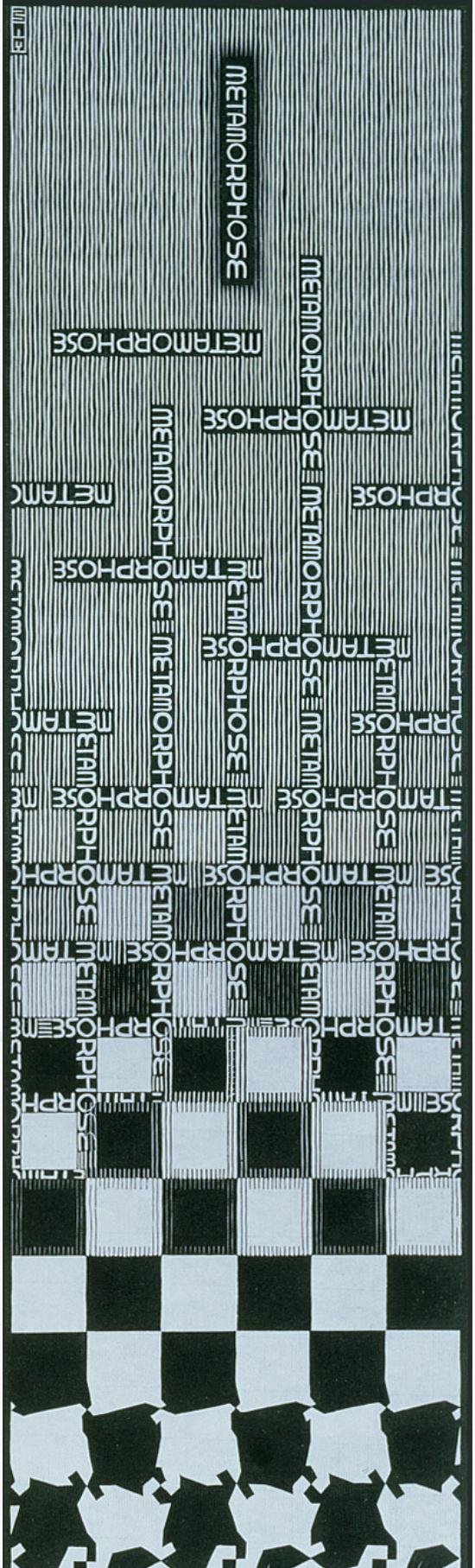
Prof. dr. M.A. Huijnen (voorzitter)
Prof. dr. C.M.A. van Ravenswaaij-Arts (Rijksuniversiteit Groningen)
Prof. dr. J.M.G. van Vugt

Voor Opa

Contents

Abbreviations

A	Adenine
AC	AmnioCentesis
ASDs	Autism Spectrum Disorders
bp	Watson and Crick base pair
bps	base pairs
C	Cytosine
ccffDNA	circulating cell-free fetal DNA
cDNA	complement DNA
(a)CGH	(array-based) Comparative Genomic Hybridization
CNV	Copy Number Variation (> 1000 bp)
CpG	Cytosine phosphate Guanine
CVS	Chorionic Villus Sampling
DNA	DesoxyriboNucleic Acid
EDTA	EthyleneDiamineTetraacetic Acid
FN	False Negative
FP	False Positive
G	Guanine
Gb	Gigabase (Billion base pairs)
GO	Gene Ontology
HPO	Human Phenotype Ontology
ID	Intellectual Disability
InDel	Insertion or Deletion variation (1-1000 bp)
IQ	Intelligence Quotient
kb	kilobase (thousand base pairs)
Mb	Megabase (Million base pairs)
ml	millilitre
MPS	Massive Parallel Sequencing
NGS	Next Generation Sequencing
NIPD	Non-Invasive Prenatal Diagnosis
NT	Nuchal Translucency
OMIM	Online Mendelian Inheritance in Man
(QF)PCR	(Quantitative Fluorescent) Polymerase Chain Reaction
pM	picoMolar
RNA	Ribo Nucleic Acid
(Z)RPKM	(Z-score adjusted) Reads Per Kilobase per Million mapped reads
SD	Standard Deviation
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant (1 bp)
SUN	Singly Unique Nucleotide
SV	Structural Variation
SVD	Singular Value Decomposition
T	Thymine
TP	True Positive
UCSC	University of California, Santa Cruz
URL	Uniform Resource Locator
WES	Whole Exome Sequencing



Chapter 1

Introduction

Based on:

J de Ligt, J A Veltman, L E L M Vissers. Point mutations as a source of *de novo* genetic disease Current Opinion in Genetics & Development 2013, 23(3):257-63

INTRODUCTION

"Why do I have blue eyes?" A typical question related to genetics usually answered with: "You inherited it from your parents". This answer is based on traditional genetics; 'The study of how living things receive traits from previous generations' [1]. We can study inherited traits, such as eye color, through the genetic information encoded by Desoxyribo Nucleic Acid (DNA). Human DNA consist of sequences of over six billion nucleotides, these nucleotides form Watson and Crick base pairs (bps) in a double helix structure. In humans, DNA is distributed over 46 chromosomes which form 23 chromosome pairs, collectively called the genome.

DNA is usually studied through a process called sequencing, once the sequence of a genomic region is known it can be compared to the sequence of other individuals. When looking for an explanation for differences in a trait, for example eye color, a comparison step is crucial, only those regions of the genome which vary in the human population are informative. When comparing the genome of two individuals only a small portion of the genome is informative, approximately 4 million bps [2,3], since most of the DNA (~99.9%) is identical between any two humans.

Genetic variation can have an effect on the physical characteristics of an individual such as eye color. Studying the DNA of individuals with either brown eyes or blue eyes for example revealed differences between the two groups in a region on chromosome 15 [4]. Within this region lies the *OCA2* gene, a gene which encodes a protein involved in melanin-based pigment formation. The level at which this gene is expressed determines the level of this pigment in the human body, higher levels of pigment result in darker skin and eye color [4].

Knowledge about the region of the genome containing the *OCA2* involved in eye color allowed researchers to study where and when the original variation occurred and how it spread through the population [4,5]. As a result several small mutations (affecting only a single base pair (bp)) have been identified in and near the *OCA2* gene which decrease the expression level in individuals with blue eyes [6,7]. Since the human ancestors had brown eyes [5] the mutation must have been introduced somewhere in our genetic history, when the first child with light brown or even blue eyes was born from parents with brown eyes. In this case a so-called *de novo* mutation occurred; a mutation was present in the DNA of the child which was not present in its parents. If a *de novo* mutation occurs during the formation of either the sperm or egg cell which is then involved in fertilization, the *de novo* mutation will be present in all cells of the child, including the germline. Those mutations present in the germline are important when studying inherited traits as only the DNA of germ cells can be transmitted to future offspring.

The example of blue eyes demonstrates how a *de novo* mutation occurred which had an effect on the phenotype and was subsequently passed on to future

***Figure 1**, Schematic overview of the different mutation types, A) Single nucleotide variant (SNV), B) Insertion or deletion event (InDel), C) Copy number variants (CNV), D) Aneuploidy, E) Inversion, F) Translocation. Italic gray text lists examples of techniques traditionally used to detect these types of mutations.*

generations. But what effect did it have on future generations and how did such a mutation occur in the first place? While one might not immediately suspect so, the mutations responsible for blue eyes have been subject to positive selection [4]. After many centuries of natural selection we can observe its outcome by studying the frequency of the mutation in the human population. The mutations responsible for blue eyes are observed most frequently in the northern parts of Europe [8]. The mutations seem to have had some beneficial effect only in this particular region of the world as they are rarely observed in populations from other regions [4]. A possible explanation for this effect is related to the fact that these are regions with less sun exposure [8].

Humans had to adapt as they migrated north from the sunny regions of Africa to the colder and relatively dark regions of northern Europe [9]. Blue eyes are a result of this adaptation process as lower amounts of pigmentation are associated to greater responsiveness to sun light [10]. A greater responsiveness to sunlight meant that individuals with these mutations required less sunlight to maintain important chemical processes such as vitamin D formation [11]. The *OCA2* mutations are an example of how *de novo* mutations can result in positive selection under certain environmental conditions.

So how did the mutations occur in the first place? The process of mutation is a natural phenomenon which occurs during the replication of DNA. When a cell divides to make two new cells the DNA of the original cell is copied. Even though the copying process is highly reliably and makes only one mistake for every 100,000 nucleotides it copies, it would accumulate to approximately 120,000 mistakes for our complete genome as over 6 billion bps (12 billion nucleotides) have to be copied [12]. However, cells also have a proof reading and error correction mechanism which corrects more than 99% of the errors introduced during the copying process [12]. DNA copying errors which are not corrected will become *de novo* mutations in a subsequent cell division cycle. The average number of mutations which occur through this process in the sperm and egg cells which are passed on to a new generation determine the so-called per generation mutation rate.

The per generation mutation rate plays an important role in human disease as was first described in 1935 [13]. It was hypothesized that certain human diseases can only be caused by *de novo* mutations because patients do not reach their reproductive age [13]. When affected individuals do not reach their reproductive age they will have no offspring which could inherit the mutation. For example Down syndrome, caused by a *de novo* duplication of chromosome 21, is associated with severe intellectual disability [14]. The additional copy of a whole chromosome has severe effects on the reproductive fitness of affected individuals from an early age [15]. A mutation with such severe effects on fitness is under strong negative selection

as it will rarely be passed on to future generations. However most mutations do not exert a phenotypic effect and some mutations, such as those in *OCA2*, may even be advantageous to an individual under certain conditions. But what kinds of mutations are exactly present and how are these detected when studying genetic diseases?

IN SEARCH OF MUTATIONS CAUSING DISEASE

Mutations occur in different types and sizes [16], and each mutation type is traditionally detected by a specialized technique (**Figure 1**). The smallest mutations affect only a single bp and are called point mutations or single nucleotide variants (SNVs) (**Figure 1a**). These mutations substitute one of the nucleotides Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) for another nucleotide, like the mutations in *OCA2* [6]. Mutations which insert or delete DNA sequences (1-1000 bps) are called insertions and deletions, or InDels collectively (**Figure 1b**). An example of an InDel is a deletion of 3 bps in the *CFTR* gene which causes cystic fibrosis [17,18]. Larger mutations (>1000 bps) that affect the amount of DNA are usually referred to as copy number variants (CNVs) and classified as either gains or losses of genomic material (**Figure 1c**), for example the 600 kb loss identified in the 17q21.31 microdeletion, or Koolen-de Vries, syndrome [19,20]. Mutations affecting a whole chromosome, thereby changing the number of chromosomes present in a cell, are called aneuploidies and occur most commonly as a trisomy (three copies of a chromosome) or a monosomy (one copy of a chromosome) (**Figure 1d**), for example trisomy 21 which causes Down syndrome [14]. Other, more complex, events can invert a piece of DNA (**Figure 1e**) or relocate a piece of DNA to another part of the genome (**Figure 1f**).

As mentioned previously, different mutation types are traditionally detected by a dedicated detection method, usually in a targeted or genome-wide manner. Targeted approaches are often more sensitive but can only be applied to small portions of the genome at a time compared to genome-wide approaches which can be applied even when no prior information is available about the genomic location. Below are two examples which illustrate how traditional techniques were applied and combined in studying the cause of severe early onset diseases with sporadic occurrence. Such disorders are typical examples of diseases where no information was known about the genomic location harboring the mutations responsible for the phenotype.

The genetic cause of CHARGE syndrome was identified after array based comparative genomic hybridization (aCGH), a genome-wide approach for CNV detection, identified a rare deletion on chromosome 8 in a patient. When the unaffected parents were tested for this mutation it was found to be absent in the DNA of both parents, the

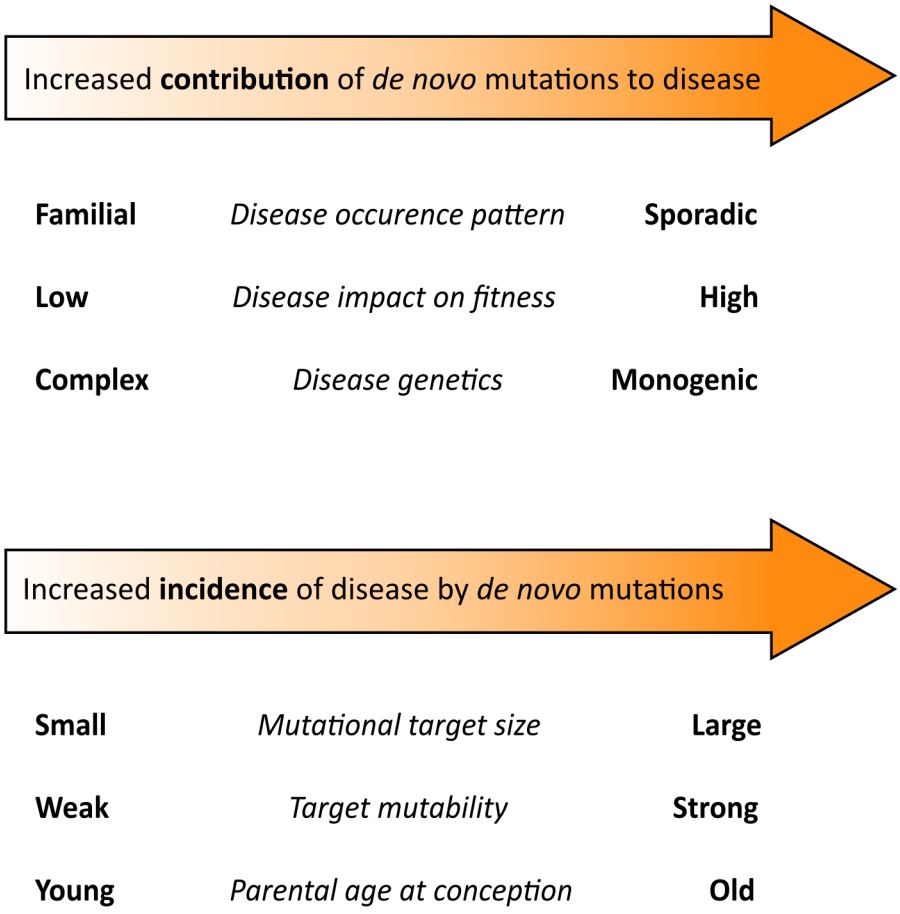


***Figure 2**, Schematic representation of the contribution of de novo mutations to genetic disease based on various disease characteristics.*

mutations was a *de novo* CNV. The genomic region identified contained the *CHD7* gene which was subsequently sequenced via Sanger sequencing to identify *de novo* SNVs and InDels in additional patients with this syndrome. Sanger sequencing of a larger cohort revealed *de novo CHD7* point mutations in a large portion (65-70%) of patients with the same phenotype, thereby establishing the cause of CHARGE syndrome [21,22]. Many other aCGH studies have shown that *de novo* CNVs occur throughout the genome and that they occur at higher frequency in individuals with neurodevelopmental disorders than in individuals without such a disorder [16]. Consequently, recurrent *de novo* microdeletions and microduplications are now recognized as a common cause of sporadic human disease [23,24].

An alternative to genome-wide approaches is the so called candidate gene approach, where researchers choose genomic regions or genes of interest based on biological knowledge. This approach is applicable to sporadic disorders but requires a sound understanding of the biological processes involved in the phenotype. An example of such a candidate gene study in neurodevelopmental disorders was the Sanger sequencing of genes involved in synaptic plasticity, an important element of human cognition. The sequencing of this sub-set of genes in a large cohort of patients with nonsyndromic intellectual disability (ID) revealed a number of *de novo* mutations affecting the *SYNGAP1* gene [25].

In the majority of patients with sporadic diseases however neither genome-wide approaches nor candidate gene strategies have been able to identify the underlying genetic defect. As a result, the cause of most rare sporadic genetic disorders remained largely undetermined. These early studies did however suggest an important role for *de novo* mutations in the sporadic occurrence of severe, early onset genetic diseases.

SYSTEMATIC GENOME-WIDE IDENTIFICATION OF *DE NOVO* MUTATIONS

In early studies, the *de novo* occurrence of mutations was determined after the mutation was identified in a patient. In other words, researchers tested the most interesting mutations for their presence in either parent. A systematic search for *de novo* mutations requires the initial experiment to be performed not only on the child but also on both parents. Sequencing of a child and its parents is also referred to as trio, or family-based sequencing. The high costs of the required laboratory experiments meant that researchers could only perform such systematic screens in a subset of the genome [25,26]. Recent technological advances, including massive parallel sequencing (MPS), have made it possible to study the whole human genome in a single experiment at base pair resolution in both the patient and its parents. MPS is also referred to as next generation sequencing (NGS) as it represents the next technological revolution in sequencing after Sanger sequencing. MPS allows researchers to study the role of *de novo* mutations in disease without a priori knowledge of the genomic regions involved.

The use of MPS technologies also resulted in datasets much larger compared to those obtained through traditional Sanger sequencing. Not only did this new technique generate many more sequence reads, (depending on the technology fragments 35 to 400 nucleotides long), the reads came from many different regions of the genome. The sheer bulk of the data combined with greater complexity required bioinformatic procedures to analyze the reads and identify genetic variation. Sequence reads had to be corrected for systematic errors, mapped to their region of origin on the genome and finally all sequencing reads had to be compared to the

reference genome to systematically identify variants. Interpreting genomic variation in a disease context required additional annotation and prioritization procedures to enable an assessment of their biological relevance.

The most complete annotations are available for those regions of the human genome which encode proteins, like the *OCA2* gene involved in eye color. All coding regions together make up approximately 1% of the genome. Mutations in these regions have been the focus of traditional research as changes in protein function can most readily be linked to changes in a patient's phenotype [27,28]. Hence many researchers chose to first focus on these regions to achieve better quality and reduce sequencing costs. The sequencing of all coding regions simultaneously is called whole exome sequencing (WES). Researchers first capture the coding regions using oligonucleotide probes after shearing the DNA into small pieces. This process of enrichment greatly increases the representation of coding regions in the final sequencing experiment.

WES was first successful in 2010 by the discovery of the genes mutated in both a rare autosomal recessive and a dominant disorder [29,30]. Since these initial studies, many groups have applied WES successfully to identify numerous other disease genes [31–33] and have confirmed the important role of *de novo* point mutations underlying rare sporadic disease [30,34–42]. Additionally, by studying *de novo* point mutations in the context of common neurodevelopmental disorders, a new paradigm emerged, accounting for the high prevalence of neurodevelopmental disorders associated with reduced fitness [**chapter 2**,43–45].

Through the systematic analysis of *de novo* mutations several factors have been identified which play a role in the occurrence of disease through *de novo* mutations. These factors are the per generation mutation rate, target mutability and target size. At the same time, the genetic architecture of the trait (familial or sporadic, monogenic or complex, normal or reduced fitness of patients) all affect the relative contribution of *de novo* mutations to the disease (**Figure 2**). Here we provide an overview of how these factors influence the occurrence and role of *de novo* mutations in disease, focusing on small mutations as these have only recently been studied.

THE HUMAN GERMLINE MUTATION RATE

An important factor which determines the frequency of germline *de novo* mutations is the per generation mutation rate. Prior to the NGS era, studies on mutation rates focused on specific genes [13,46,47]. For instance, the evaluation of 20 loci involved in Mendelian diseases yielded a direct estimate of 1.8×10^{-8} point mutations per nucleotide per generation. This indicated that single nucleotide substitutions were 25 times more common than all other mutations [47–49]. In 2010, a first systematic

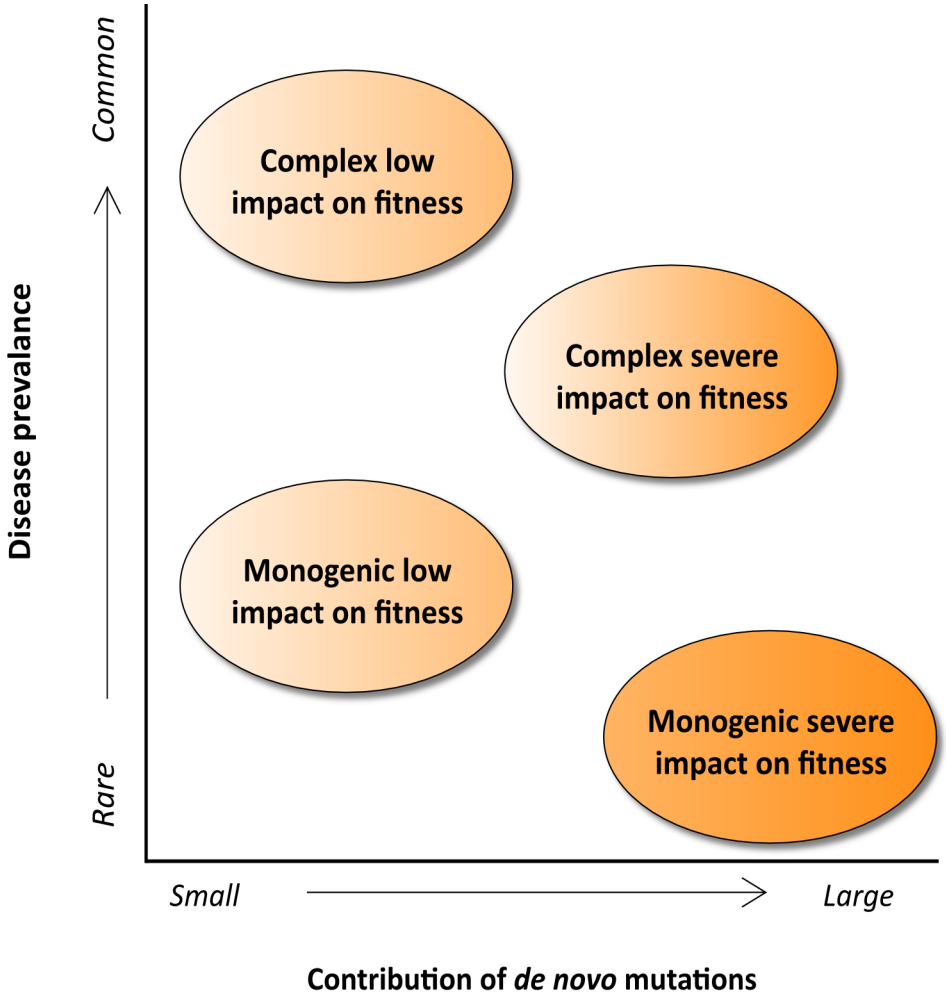


***Figure 3**, Schematic representation of the relation between the contribution of de novo mutations to disease type and prevalence. Orange shading reflects the impact of de novo mutations, with darker shading indicating a more prominent role for de novo mutations in diseases with these characteristics.*

study investigated the occurrence of *de novo* point mutations in patients with neurodevelopmental disorders by resequencing >400 candidate disease genes. This study suggested an equal number of *de novo* mutations in patients and controls but showed an excess of deleterious mutations in patients [43].

The availability of direct measurements of large amounts of DNA sequence has led to a more accurate determination of the human germline mutation rate [50,51] which has proven to be close to its early estimates [47,52]. The current best estimate

of the human germline point mutation rate is 1.18 x 10^{-8} per nucleotide, which corresponds to 74 *de novo* point mutations throughout the genome per generation [16]. In contrast, the larger mutation types have lower germline mutation rates, 4×10^{-10} for InDels [53] whereas large (>100kb) CNVs occur approximately in only 1 out of every 50 individuals [54].

Genomic positions that frequently mutate, so called mutation hotspots, were found to have only a minor contribution to overall human mutation rates [51]. Interestingly, Conrad and colleagues [51] showed substantial variation in the germline mutation rate, not only between families, but also within a family. Also, several WES studies in autism have reported an increased number of *de novo* point mutations as well as an excess of deleterious *de novo* point mutations in patients compared to controls [55,56]. These observations may, in part, be explained by a recent large-scale parent-offspring study that was able to demonstrate that the number of *de novo* germline mutations increases with advancing paternal age [57] (see also section on the impact of parental age on the occurrence of *de novo* mutations). Comparison of various large-scale sequencing studies [44,55,58,59] indicates that of all *de novo* point mutations, an average of 1.19 mutations per generation affect the coding sequence. A large fraction (~78%) of the mutations affecting the coding sequence is predicted to alter the protein sequence [16], possibly altering protein function.

The per generation germline mutation rates discussed here provide insights into the average number of mutations which occur but do not address the underlying mechanisms. The mutational mechanisms do however determine the likelihood that a certain region is mutated, or target mutability.

TARGET MUTABILITY

In principle DNA copying errors, and therefore mutations, occur randomly. However, the probability of mutation at a given site is not uniform across the genome [60]. Regional mutation rates are subject to a variety of genomic characteristics [61] and influenced by external factors such as parental age [57,62,63] (see the parental age section). Large differences in mutational rate are particularly evident for structural variation (SV) for which mutation rates have been observed to vary between 10^{-4} and 10^{-6} mutations per base per generation [64]. Some of these differences can be explained by SV hot spots, where recurrent mutations are mediated by nonallelic homologous recombination (NAHR) between tandem segmental duplications [65,66]. Regional rates of nucleotide substitution are also variable [61] which depends on a large number of factors [60]. In contrast to the SV hot spots (predominantly driven by meiotic recombination) nucleotide substitution can occur by a variety of mechanisms, and the mutation rate is influenced to a much greater extent by mitotic mechanisms [62].

***Table 1**, The effect of the mutational target size and de novo point mutations on the frequency of disease*

Estimated mutational target size	Example genetic disorder name (Gene)	Observed disease frequency (x/100,000)	Expected *de novo* mutation frequency for target (x/100,000)
1 nucleotide (bp)	New ID syndrome (PACS1)	<0.001 a	0.0019 d
11 nucleotides (bps)	Schinzel-Giedion (SETBP1)	<0.01 a	0.021 d
~300 nucleotides (bps) *	Noonan (>6 genes)	50 b	0.58 d
1 gene	CHARGE (CDH7)	0.14 b	0.26 e,f
~500 genes	Severe ID	300-500 c	601 f,g

** The mutational target size for Noonan syndrome was adjusted from the original publication to better reflect the gain of function mechanism of this disease, to date 92 codons have been implicated in Noonan syndrome [Uniprot, accessed 01-08-2013] assuming that 2 bases in each codon can cause the disease when mutated and that current knowledge explains approximately 60% of cases [76] results in an estimated target size of 307 base pairs to explain 100% of the cases.*

Of note, observed frequencies exceeding the expected frequency due to de novo mutations may indicate familial occurrence, clonal expansion in spermatogenesis and/or other genetic factors playing a role, whereas observed frequencies lower than expected may suggest embryonic lethality.

a. Based on the number of currently published cases.
b. Derived from; rare disease frequency report of ORPHANET [101].
c. Estimated mutation target size as reported by Leonard and Wen [15].
d. An average non-synonymous germline mutation rate of 0.58 per generation [16].
e. An average truncating germline mutation rate of 0.055 per generation [16].
f. The expected frequency of a disease in the population was based on the following assumptions:
  - *an average coding gene size of 1419 base pairs*
  - *de novo point mutations occur randomly in the coding sequence*

g. To estimate the number of mutations functionally affecting genes, a causal mutation rate of 0.252 per generation was used, based on 0.055 + (0.58 × 0.34); the average truncating mutation rate plus 34% [102] of the average non-synonymous germline mutation rate.

Germline *de novo* point mutations have been found to show nonrandom occurrence in the genome. Compared to a random mutation model *de novo* mutations are spaced more closely than expected [60]. Closely spaced *de novo* mutations, called mutation pairs, were observed to have a single parent of origin, consistent with mutations arising in a single replication event [67]. Clusters of two or more *de novo* mutations within 100 kb of each other occur approximately once per generation [60]. The occurrence of such clusters could be explained by compound mutation or by *de novo* nucleotide substitutions that occur during allelic gene conversion events [68,69].

Numerous genomic features have been found to influence site mutability, of which the most significant are DNase hypersensitivity, GC content, nucleosome occupancy, recombination rate, trinucleotide sequence content and simple repeats surrounding the site [60]. Most *de novo* mutations constitute transitions (purine ↔ purine and/or pyrimidine ↔ pyrimidine) rather than transversions (purine ↔ pyrimidine). Cytosine phosphate Guanine (CpG)-rich regions are enriched for *de novo* mutations as indicated by an elevated mutation rate of 1.5×10^{-8} per nucleotide [13,48,49,56], occurring most often on CpG dinucleotides [60]. Germline methylation further influences the substitution rate at these CpG sites [70]. The elevated mutation rate in GC-rich regions may have consequences for human disease as most disease-causing mutations identified to date are located in coding sequences, which are GC rich.

THE ROLE OF GERMLINE *DE NOVO* MUTATIONS IN RARE GENETIC DISEASE

The search for disease-causing *de novo* mutations in rare sporadic syndromes has been accelerated considerably by WES. The cause of Kabuki syndrome (*MLL2*) [34] was described, and the gene for Schinzel Gieion Syndrome was identified the same year by *de novo* mutations in *SETBP1* for 12 of 13 patients [30]. Many successes followed, including, Hajdu-Cheney syndrome (*NOTCH2*) [35,36], KBG syndrome (*ANKRD11*) [37], Baraitser-Winter syndrome (*ACTB* and *ACTG1*) [38], Coffin-Siris syndrome (components of the SWI/SNF complex) [39,40], and Cantù syndrome (*ABCC9*) [41,42]. These studies have shown that sporadic diseases are now amenable to genetic disease research. An important success factor in these studies was careful selection of a homogeneous groups for analysis by thorough phenotypic characterization of patients. The proper grouping of patients leads to greater similarity in the underlying disease mechanism and thereby the underlying genetics. For rare Mendelian disorders, exome sequencing the DNA of three to four well-phenotyped patients combined with an analysis to identify *de novo* point mutations in the same gene, or gene(s) in the same pathway, is sufficient to discover the genetic cause of disease [39,40,71,72]. Similarly, a WES study of

two unrelated patients with ID and a striking facial resemblance suggestive of a hitherto unappreciated syndrome identified the exact same *de novo* point mutation in *PACS1* in both of them, underscoring the need for detailed phenotyping [73]. To facilitate the identification of such syndromes, data sharing at international level for both genotypes as well as phenotypes is of utmost importance (see also section below on *de novo* point mutations in common genetic disease and **chapter 6**).
These studies also provided insight in the interactions between fitness, severity of disease and the role of *de novo* mutations. In KBG syndrome the majority of sporadic patients were shown to have a *de novo* mutation in *ANKRD11*, however one mutation segregated with disease in a family [37]. This family was reported to have a less severe ID phenotype, which potentially decreased the selective pressure on this mutation. Similarly, severe late onset genetic disorders, such as Alzheimer Disease, have little to no effect on reproductive fitness, allowing pathogenic mutations to spread through the population [74], thereby decreasing the relative contribution of *de novo* mutations to disease (**Figures 2 & 3**). It has to be noted that a comprehensive analysis of *de novo* mutations in late-onset neurodegenerative disorders is hampered by the fact that patients with these diseases usually do not have surviving parents who can supply DNA samples for inheritance analysis [75].

MUTATIONAL TARGET SIZE AND DISEASE FREQUENCY

The frequency of *de novo* mutations causing a certain disease reflects the number and size of the genomic regions that are involved. The aggregate of these regions is also known as the mutational target. The chance that random *de novo* mutations affect a target is largely determined by the targets size (**Figure 2**). For example in *SETBP1*, the gene involved in Schinzel-Giedion syndrome, all pathogenic *de novo* point mutations were contained within an 11 base pair stretch [30]. This strong clustering of mutations results in a very small mutational target, which predicts that this syndrome should be extremely rare. Its frequency is determined by the chance that one of the approximately 74 *de novo* germline point mutations affects this 11 base pair locus. The chance of hitting the target should be even lower when a disease is caused by mutation of only one specific base, as was recently described for *PACS1* [73]. In the case of *PASC1* only two patients worldwide have been recognized with this syndrome so far, although other patients must surely exist [73].
In contrast to rare genetic disorders, the central hypothesis for common diseases has been that they can be caused by common genetic variants which are detectable by adequately powered genome-wide association studies (GWAS). *De novo* mutations affecting a large mutational target of hundreds, or even thousand(s), of genes could collectively cause a common genetic disease (**Figure 3**).
Random modeling of *de novo* mutations for well-known monogenic diseases, with

and without locus-heterogeneity, indicates that the frequency of diseases caused by *de novo* point mutations indeed mostly reflects the mutational target size (**Table 1**). Notably, it may be hypothesized that the mutational target size will provide insight into the complexity of the underlying process. More specifically, the high prevalence of neurodevelopmental disorders might reflect the complexity of the central nervous system, which is regulated by many different pathways. Hence, every gene that regulates, or is a part of, these pathways may contribute to disease when mutated.

Assuming target size as the main determinant, we can see that a large discrepancy exists for Noonan syndrome between the expected and observed disease frequency (**Table 1**). This may in part be due to the familial occurrence of this disorder [76] but familial transmission of Noonan is relatively rare, and can certainly be only a minor factor in its high prevalence. A more likely cause is clonal expansion of Noonan syndrome gene mutations in the spermatogonia [77]. This spermatogonial expansion correlates strongly with paternal age [26,77] and is discussed in more detail in the parental age section.

Generally speaking, gain-of-function or dominant negative disease mechanisms create smaller mutational targets than do loss of function mechanisms. For example, all mutations involved in Schinzel-Giedion syndrome were confined to an 11 base pair stretch which may be indicative for a gain-of-function mechanism [30]. This notion is further supported by the identification of other clinical entities in which *SETBP1* plays a role; deletions including this gene contribute to the 18q contiguous gene deletion syndrome [78], whereas truncating mutations have been implicated as a cause of autism spectrum disorders (ASDs) and nonsyndromic intellectual disability [79,80]. These genotype-phenotype correlations based on the underlying pathophysiological mechanism are well-known in other diseases [81] and will have a large effect on disease frequency.

DE NOVO MUTATIONS IN COMMON GENETIC DISEASE

Prior to the introduction of NGS, the large degree of locus-heterogeneity complicated a comprehensive study of rare and *de novo* point mutations in common disease, such as for instance ID, and ASD. In earlier studies genome-wide approaches such as GWAS and aCGH have implicated many loci in such complex disorders, but these often required additional follow-up to elucidate the underlying cellular mechanism [65,82]. Using family-based WES approaches, a number of studies have recently supported the hypothesis that *de novo* mutations indeed affect a wide variety of genes and that these events collectively play an important role in common neurodevelopmental diseases such as ID [**chapter 2 & 3**,80], ASDs [44,58,59,79] and schizophrenia [43,45,83].

The first systematic family-based WES study of 10 patients with severe ID identified possibly causal *de novo* mutations in 6 out of 10 patients and thereby established a *de novo* paradigm for ID [**chapter 2**]. Two larger studies subsequently found causal deleterious *de novo* point mutations in 13-36% of sporadic patients with severe ID [**chapter 3**, 80]. A large number of other patients carried *de novo* mutations in genes with a role in brain development or function but which, had not previously been found mutated in ID. If all these candidates genes would be relevant for ID, the percentage of patients with causal *de novo* point mutations may rise as high as 32-60% [**chapter 3**, 80]. Similarly, based on initial studies in known disease genes, *de novo* point mutations seem to contribute to ASDs and schizophrenia [43,45,58,59,79,83] albeit with a lower impact than in severe ID. A WES study of sporadic schizophrenia patients did not reveal mutations in known schizophrenia genes but yielded candidate genes with presumably causal *de novo* point mutations in 17% of cases [45,83]. Three other large-scale studies used WES to study the role of *de novo* mutations in ASDs [58,59,79]. The studies identified 2-4% causal *de novo* point mutations in known and novel ASDs associated genes. Although the exact number of candidate genes for ASDs is unclear due to the different criteria used for candidate classification, 20 to 50 novel ASDs genes may be defined. Attributing causality to the candidate genes would increase the contribution of *de novo* point mutations in ASDs to approximately 17%.

The results for ID, ASD and schizophrenia collectively underscore the importance of *de novo* mutations for common genetic disorders as well as the need for follow-up studies to establish the relevance of (*de novo*) mutations in candidate disease genes. A complicating factor for interpretation is that many *de novo* mutations in candidate genes involve missense mutations for which it is difficult to establish pathogenicity and specific disease mechanisms without additional functional follow-up of each mutation individually. Current software prediction tools, such as SIFT [84], PolyPhen-2 [85], Condel [86] and MutationTaster [87], are not sufficiently robust to provide a reliable and unambiguous prediction on mutation impact [88].

For monogenic disorders proving causality has been less of a challenge as researchers can rely on the detection of mutations in the same gene in patients with overlapping phenotypes [30,34–42]. For diseases with a larger degree of locus heterogeneity, efforts have been put in targeted re-sequencing of candidate genes to increase the chance of finding additional mutations [**chapter 3**,79,83,89]. The success of this approach relies heavily on the patient cohort selected, and thus on (deep-) phenotyping of patients using standardized nomenclature such as the human phenotype ontology (HPO) [90].

Clinically well-defined patient cohorts also facilitate reverse-phenotyping, where phenotypes are (re-)defined based on genetic information that has become available.

For instance, we recently found two patients in different studies with severe ID and a *de novo* mutation in *DYNC1H1* [**chapter 2 & 3**]. Re-evaluation of the phenotypes of these patients revealed a neuronal migration defect and other clinical similarities, thereby potentially defining a new syndrome [91]. There are likely many more "recognizable clinical entities within a common disease", which predicts that many novel syndromes will be delineated in the near future. The availability of precise and accurate phenotypes is of special importance given the fact that recurrence of certain *de novo* mutations may only be identified by examining patients from across the globe, especially for small mutational targets. Disease (gene) discovery may therefore benefit from ongoing efforts to sustain databases that contain both detailed phenotypes as well as genotypes. It is important that these databases can also contain different forms of genetic variation [**chapter 4**,92,93]. In addition, functional studies remain essential In order to establish the pathogenic nature of *de novo* mutations and to understand the pathogenic mechanisms involved. An example is the mechanism by which *de novo* missense mutations of *ABCC9* cause Cantù syndrome [42].

THE IMPACT OF PARENTAL AGE ON THE OCCURRENCE OF *DE NOVO* MUTATIONS

Down syndrome, is one of the most common and well known disorders associated with ID and parental age [94]. John Langdon Down defined the syndrome in 1866. In 1959 Jérôme Lejeune was able to count the chromosomes of healthy and affected individuals, referred to as karyotyping, with the use of a revolutionary technique in tissue culturing. He observed that all affected individuals had three copies, called a trisomy, of chromosome 21 rather than two copies as observed in unaffected individuals. This experiment identified the genetic cause underlying Down syndrome and provided a first molecular diagnosis for ID. More detailed studies on whole chromosome abnormalities, aneuploidies, found a strong bias in parental origin, events being mostly of maternal origin and subsequently the occurrence rate was found to be associated with maternal age [94], these observations have led to screening for trisomies in pregnancies after the age of 36 years in the Netherlands [95]. More recently, it has become apparent that the identification of trisomy 21 is also possible in a non-invasive manner using MPS [96,97, **chapter 5**]. This technique relies on the presence of fetal DNA material in the blood circulation of the pregnant women [96]. The quantification of the maternal and fetal DNA simultaneously allows researchers to measure relative abundances of different chromosomes [96], further discussed in **chapter 5**.

In contrast to aneuploidies mostly being of maternal origin, recent large-scale sequencing studies have shown that approximately 80% of *de novo* point mutations occur on the paternally-derived chromosome [44,57]. Further, the majority of *de novo*

CNVs appear to be of paternal origin [63]. In general, a strong correlation between paternal age and the number of *de novo* point mutations has now been established [57]. For the common male reproductive period (20-40 years, with an average age of 29.7) the number of *de novo* mutations in the offspring is estimated to increase by ~4% each year [57]. This corresponds to two additional *de novo* mutations on an annual basis and a doubling of the total number of *de novo* mutations every 16.5 years [57]. As *de novo* point mutations occur more often in Guanine and Cytosine (GC) rich regions (see section on target mutability), the increase in the number of mutations in coding regions might be even higher. Assuming a two-fold increase in *de novo* mutations with advanced paternal age would result in a two-fold increase in genetic diseases caused by *de novo* mutations. This elevated risk may be even stronger if other genetic as well as gestational and environmental factors are taken into account [98].

To estimate the relative increased risk of a genetic disorder with advanced paternal age the contribution of mutation and of inheritance must be considered. In the case of Schinzel-Giedion, caused by *de novo* missense mutations in a mutational target of 11 base pairs, we can assume that all cases reflect *de novo* events. Using the non-synonymous germline mutation rate of 0.58 per generation [16], and assuming that all *de novo* point mutations leading to this disease are of paternal origin, a two-fold increase in the number of *de novo* mutations will be equivalent to a two-fold increase in risk of the disorder from 0.021 to 0.042 in every 100,000 live births. More complex disorders, such as ASDs or mild ID, are considered only partly genetic and only a portion of the genetic contribution is explained by *de novo* mutations. Hence, a two-fold increase in *de novo* mutations may only slightly raise the occurrence of these diseases. Nonetheless, it is tempting to speculate that the apparent increase in incidence and prevalence of neurodevelopmental conditions such as ASDs could partly be due to the accumulation of mutations in the population as an effect of advancing paternal age [99]. There is an important class of disorders for which the paternal age effect is much stronger. This includes *FGFR2* mutations for Apert Syndrome and *FGFR3* mutations for achondroplasia and thanatophoric dysplasia [100]. For these so called “paternal age effect” disorders, *de novo* mutations are positively selected and expand clonally in normal testes through a process similar to oncogenic expansion [26]. Positive germline selection has also been demonstrated for Noonan syndrome (**Table 1**) [77]. It is important to note that all mutations that appear to show this phenomenon of testicular selection by clonal expansion are implied in oncogenic pathways, and have gain of function effects. It may be expected that further examples exist possibly also including neurodevelopmental disorders [26].

GENERAL CONCLUSIONS

New technologies have enabled the genome-wide evaluation of genetic variation and this has established that *de novo* germline mutations are a major cause of monogenic diseases with a severe impact on reproductive fitness in both rare and common disorders. The clinical interpretation of rare *de novo* missense variants remains a challenge as and requires a combination of deep-phenotyping, testing for recurrently mutated genes, functional approaches and international data-sharing to further our understanding of genetic disease [**chapter 6**].

REFERENCES

1. Genetics [Forbes AA, Krimmel BA.: Evolution Is Change in the Inherited Traits of a Population through Successive Generations. Nature Education Knowledge 2010, 3:6]
2. Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, Egholm M, et al.: A map of human genome variation from population-scale sequencing. Nature 2010, 467:1061–73.
3. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al.: Origins and functional impact of copy number variation in the human genome. Nature 2010, 464:704–12.
4. Sturm RA, Frudakis TN: Eye colour: portals into pigmentation genes and ancestry. Trends in genetics : TIG 2004, 20:327–32.
5. Meyer WK, Zhang S, Hayakawa S, Imai H, Przeworski M: The convergent evolution of blue iris pigmentation in primates took distinct molecular paths. American journal of physical anthropology 2013, 151:398–407.
6. Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA: A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. American journal of human genetics 2007, 80:241–52.
7. Sturm RA, Duffy DL, Zhao ZZ, Leite FPN, Stark MS, Hayward NK, Martin NG, Montgomery GW: A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. American journal of human genetics 2008, 82:424–31.
8. Coon CS: The Races of Europe. 1939.
9. Cavalli-Sforza LL, Feldman MW: The application of molecular genetic approaches to the study of human evolution. Nature genetics 2003, 33 Suppl:266–75.
10. Tomany SC, Klein R, Klein BEK: The relationship between iris color, hair color, and skin sun sensitivity and the 10 year incidence of age-related maculopathy: the Beaver Dam Eye Study. Ophthalmology 2003, 110:1526–33.
11. Premkumar M, Sable T, Dhanwal D, Dewan R: Vitamin D homeostasis, bone mineral metabolism, and seasonal affective disorder during 1 year of Antarctic residence. Archives of osteoporosis 2013, 8:129.
12. Pray, L. (2008) DNA replication and causes of mutation. Nature Education 1(1)

13. Haldane JBS: The rate of spontaneous mutation of a human gene. Journal of Genetics 1935, 31:317–326.
14. Korbel JO, Tirosh-Wagner T, Urban AE, Chen X-N, Kasowski M, Dai L, Grubert F, Erdman C, Gao MC, Lange K, et al.: The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies. Proceedings of the National Academy of Sciences of the United States of America 2009, 106:12031–6.
15. Leonard H, Wen X: The epidemiology of mental retardation: challenges and opportunities in the new millennium. Mental retardation and developmental disabilities research reviews 2002, 8:117–34.
16. Veltman JA, Brunner HG: *De novo* mutations in human genetic disease. Nature reviews. Genetics 2012, 13:565–75.
17. Collins FS, Drumm ML, Cole JL, Lockwood WK, Vande Woude GF, Iannuzzi MC: Construction of a general human chromosome jumping library, with application to cystic fibrosis. Science (New York, N.Y.) 1987, 235:1046–9.
18. Mullaney JM, Mills RE, Pittard WS, Devine SE: Small insertions and deletions (INDELs) in human genomes. Human molecular genetics 2010, 19:R131–6.
19. Koolen DA, Vissers LELM, Pfundt R, De Leeuw N, Knight SJL, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, et al.: A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nature genetics 2006, 38:999–1001.
20. Koolen DA, Kramer JM, Neveling K, Nillesen WM, Moore-Barton HL, Elmslie F V, Toutain A, Amiel J, Malan V, Tsai AC-H, et al.: Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nature genetics 2012, 44:639–41.
21. Vissers LELM, Van Ravenswaaij CM a, Admiraal R, Hurst J a, De Vries BBA, Janssen IM, Van der Vliet WA, Huys EHLPG, De Jong PJ, Hamel BCJ, et al.: Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. Nature genetics 2004, 36:955–7.
22. CHARGE 2 [Lalani SR, Hefner MA, Belmont JW, Davenport SLH: CHARGE syndrome. In Gene Reviews tm. Edited by Pagon RA, Bird TD, Dolan CR, Stephens K, Adam MP. 2012. (PMID: 20301296)
23. Zhang F, Gu W, Hurles ME, Lupski JR: Copy number variation in human health, disease, and evolution. Annual review of genomics and human genetics 2009, 10:451–81.
24. Girirajan S, Campbell CD, Eichler EE: Human copy number variation and complex genetic disease. Annual review of genetics 2011, 45:203–26.
25. Hamdan FF, Gauthier J, Spiegelman D, Noreau A, Yang Y, Pellerin S, Dobrzeniecka S, Côté M, Perreau-Linck E, Perreault-Linck E, et al.: Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. The New England journal of medicine 2009, 360:599–605.
26. Goriely A, Wilkie AOM: Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. American journal of human genetics 2012, 90:175–200.
27. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, et al.: Clinical assessment incorporating a personal genome.

Lancet 2010, 375:1525–35.

28. Gonzaga-Jauregui C, Lupski JR, Gibbs R a: Human genome sequencing in health and disease. Annual review of medicine 2012, 63:35–61.
29. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al.: Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 2010, 42:30–5.
30. Hoischen A, Van Bon BWM, Gilissen C, Arts P, Van Lier B, Steehouwer M, De Vries P, De Reuver R, Wieskamp N, Mortier G, et al.: *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. Nature genetics 2010, 42:483–5.
31. Gilissen C, Hoischen A, Brunner HG, Veltman JA: Disease gene identification strategies for exome sequencing. European journal of human genetics : EJHG 2012, 20:490–7.
32. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: Exome sequencing as a tool for Mendelian disease gene discovery. Nature reviews. Genetics 2011, 12:745–55.
33. Gilissen C, Hoischen A, Brunner HG, Veltman JA: Unlocking Mendelian disease using exome sequencing. Genome biology 2011, 12:228.
34. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, et al.: Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics 2010, 42:790–3.
35. Simpson MA, Irving MD, Asilmaz E, Gray MJ, Dafou D, Elmslie F V, Mansour S, Holder SE, Brain CE, Burton BK, et al.: Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. Nature genetics 2011, 43:303–5.
36. Isidor B, Lindenbaum P, Pichon O, Bézieau S, Dina C, Jacquemont S, Martin-Coignard D, Thauvin-Robinet C, Le Merrer M, Mandel J-L, et al.: Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. Nature genetics 2011, 43:306–8.
37. Sirmaci A, Spiliopoulos M, Brancati F, Powell E, Duman D, Abrams A, Bademci G, Agolini E, Guo S, Konuk B, et al.: Mutations in ANKRD11 cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia. American journal of human genetics 2011, 89:289–94.
38. Rivière J-B, Van Bon BWM, Hoischen A, Kholmanskikh SS, O'Roak BJ, Gilissen C, Gijsen S, Sullivan CT, Christian SL, Abdul-Rahman OA, et al.: *De novo* mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. Nature genetics 2012, 44:440–4, S1–2.
39. Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, Kaname T, Naritomi K, Kawame H, Wakui K, et al.: Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nature genetics 2012, 44:376–8.
40. Santen GWE, Aten E, Sun Y, Almomani R, Gilissen C, Nielsen M, Kant SG, Snoeck IN, Peeters EAJ, Hilhorst-Hofstee Y, et al.: Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. Nature genetics 2012, 44:379–80.
41. Van Bon BWM, Gilissen C, Grange DK, Hennekam RCM, Kayserili H, Engels H, Reutter H, Ostergaard JR, Morava E, Tsiakas K, et al.: Cantú syndrome is caused by mutations in ABCC9. American journal of human genetics 2012, 90:1094–101.

42. Harakalova M, Van Harssel JJT, Terhal PA, Van Lieshout S, Duran K, Renkens I, Amor DJ, Wilson LC, Kirk EP, Turner CLS, et al.: Dominant missense mutations in ABCC9 cause Cantú syndrome. Nature genetics 2012, 44:793–6.
43. Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, Griffing AR, Côté M, Henrion E, Spiegelman D, Tarabeux J, et al.: Direct measure of the *de novo* mutation rate in autism and schizophrenia cohorts. American journal of human genetics 2010, 87:316–24.
44. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, et al.: Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. Nature genetics 2011, 43:585–9.
45. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, Gogos JA, Karayiorgou M: Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. Nature genetics 2011, 43:864–8.
46. Kondrashov AS, Crow JF: A molecular approach to estimating the human deleterious mutation rate. Human mutation 1993, 2:229–34.
47. Kondrashov AS: Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Human mutation 2003, 21:12–27.
48. Arnheim N, Calabrese P: Understanding what determines the frequency and pattern of human germline mutations. Nature reviews. Genetics 2009, 10:478–88.
49. Hodgkinson A, Eyre-Walker A: Variation in the mutation rate across mammalian genomes. Nature reviews. Genetics 2011, 12:756–66.
50. Roach JC, Glusman G, Smit AF a, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al.: Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science (New York, N.Y.) 2010, 328:636–9.
51. Conrad DF, Keebler JEM, DePristo M a, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella K V, et al.: Variation in genome-wide mutation rates within and between human families. Nature genetics 2011, 43:712–4.
52. Vogel F, Rathenberg R: Spontaneous mutation in man. Advances in human genetics 1975, 5:223–318.
53. Lynch M: Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences of the United States of America 2010, 107:961–8.
54. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE: *De novo* rates and selection of large copy number variation. Genome research 2010, 20:1469–81.
55. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee Y-H, Narzisi G, Leotta A, et al.: *De novo* gene disruptions in children on the autistic spectrum. Neuron 2012, 74:285–99.
56. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, Dionne-Laporte A, Spiegelman D, Henrion E, Diallo O, et al.: Increased exonic *de novo* mutation rate in individuals with schizophrenia. Nature genetics 2011, 43:860–3.
57. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir AA, Wong WSW, et al.: Rate of *de novo* mutations and the importance of father's age to disease risk. Nature 2012, 488:471–5.

58. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin C-F, Stevens C, Wang L-S, Makarov V, et al.: Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. Nature 2012, 485:242–5.

59. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, et al.: *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 2012, 485:237–41.

60. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al.: Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. Cell 2012, 151:1431–42.

61. Ellegren H, Smith NGC, Webster MT: Mutation rate variation in the mammalian genome. Current opinion in genetics & development 2003, 13:562–8.

62. Crow JF: The origins, patterns and implications of human spontaneous mutation. Nature reviews. Genetics 2000, 1:40–7.

63. Hehir-Kwa JY, Rodríguez-Santiago B, Vissers LE, De Leeuw N, Pfundt R, Buitelaar JK, Pérez-Jurado LA, Veltman JA: *De novo* copy number variants associated with intellectual disability have a paternal origin and age bias. Journal of medical genetics 2011, 48:776–8.

64. Lupski JR: Genomic rearrangements and sporadic disease. Nature genetics 2007, 39:S43–7.

65. Malhotra D, Sebat J: CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell 2012, 148:1223–41.

66. Lupski JR: Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends in genetics : TIG 1998, 14:417–22.

67. Wang J, Gonzalez KD, Scaringe WA, Tsai K, Liu N, Gu D, Li W, Hill KA, Sommer SS: Evidence for mutation showers. Proceedings of the National Academy of Sciences of the United States of America 2007, 104:8403–8.

68. Hurles M: Are 100,000 "SNPs" useless? Science (New York, N.Y.) 2002, 298:1509; author reply 1509.

69. Rattray AJ, Shafer BK, McGill CB, Strathern JN: The roles of REV3 and RAD57 in double-strand-break-repair-induced mutagenesis of Saccharomyces cerevisiae. Genetics 2002, 162:1063–77.

70. Mugal CF, Ellegren H: Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. Genome biology 2011, 12:R58.

71. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al.: Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009, 461:272–6.

72. Hoischen A, Van Bon BWM, Rodríguez-Santiago B, Gilissen C, Vissers LELM, De Vries P, Janssen I, Van Lier B, Hastings R, Smithson SF, et al.: *De novo* nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nature genetics 2011, doi:10.1038/ng.868.

73. Schuurs-Hoeijmakers JHM, Oh EC, Vissers LELM, Swinkels MEM, Gilissen C, Willemsen M a, Holvoet M, Steehouwer M, Veltman J a, De Vries BB a, et al.: Recurrent *De novo*

Mutations in PACS1 Cause Defective Cranial-Neural-Crest Migration and Define a Recognizable Intellectual-Disability Syndrome. American journal of human genetics 2012, doi:10.1016/j.ajhg.2012.10.013.

74. Tanzi RE: The genetics of Alzheimer disease. Cold Spring Harbor perspectives in medicine 2012, 2.
75. Pamphlett R, Morahan JM, Yu B: Using case-parent trios to look for rare *de novo* genetic variants in adult-onset neurodegenerative diseases. Journal of neuroscience methods 2011, 197:297–301.
76. Roberts AE, Allanson JE, Tartaglia M, Gelb BD: Noonan syndrome. Lancet 2013, 381:333–42.
77. Yoon S-R, Choi S-K, Eboreime J, Gelb BD, Calabrese P, Arnheim N: Age-Dependent Germline Mosaicism of the Most Common Noonan Syndrome Mutation Shows the Signature of Germline Selection. American journal of human genetics 2013, doi:10.1016/j.ajhg.2013.05.001.
78. Buysse K, Menten B, Oostra A, Tavernier S, Mortier GR, Speleman F: Delineation of a critical region on chromosome 18 for the del(18)(q12.2q21.1) syndrome. American journal of medical genetics. Part A 2008, 146A:1330–4.
79. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, et al.: Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. Nature 2012, 485:246–50.
80. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, et al.: Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 2012, 6736:1674–82.
81. Halliday G, Bigio EH, Cairns NJ, Neumann M, Mackenzie IR a, Mann DM a: Mechanisms of disease in frontotemporal lobar degeneration: gain of function versus loss of function effects. Acta neuropathologica 2012, 124:373–82.
82. Reich DE, Lander ES: On the allelic spectrum of human disease. Trends in genetics : TIG 2001, 17:502–10.
83. Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y, Levy S, Gogos J a, Karayiorgou M: *De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nature genetics 2012, doi:10.1038/ng.2446.
84. Kumar P, Henikoff S, Ng PC: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 2009, 4:1073–81.
85. Adzhubei I a, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: A method and server for predicting damaging missense mutations. Nature methods 2010, 7:248–9.
86. González-Pérez A, López-Bigas N: Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. American journal of human genetics 2011, 88:440–9.
87. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D: MutationTaster evaluates disease-causing potential of sequence alterations. Nature methods 2010, 7:575–6.

88. Gray VE, Kukurba KR, Kumar S: Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. Bioinformatics (Oxford, England) 2012, 28:2093–6.

89. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, et al.: Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science (New York, N.Y.) 2012, 338:1619–22.

90. Robinson PN, Mundlos S: The human phenotype ontology. Clinical genetics 2010, 77:525–34.

91. Willemsen MH, Vissers LEL, Willemsen M a a P, Van Bon BWM, Kroes T, De Ligt J, De Vries BB, Schoots J, Lugtenberg D, Hamel BCJ, et al.: Mutations in DYNC1H1 cause severe intellectual disability with neuronal migration defects. Journal of medical genetics 2012, 49:179–83.

92. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A: Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics (Oxford, England) 2012, 28:2711–8.

93. Ritz A, Bashir A, Raphael BJ: Structural variation analysis with strobe reads. Bioinformatics (Oxford, England) 2010, 26:1291–8.

94. Morris JK, Mutton DE, Alberman E: Revised estimates of the maternal age specific live birth prevalence of Down's syndrome. Journal of medical screening 2002, 9:2–6.

95. Morris JK, Waters JJ, De Souza E: The population impact of screening for Down syndrome: audit of 19 326 invasive diagnostic tests in England and Wales in 2008. Prenatal diagnosis 2012, 32:596–601.

96. Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FMF, Zheng YW, Leung TY, Lau TK, Cantor CR, et al.: Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Science translational medicine 2010, 2:61ra91.

97. Fan HC, Quake SR: Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. PloS one 2010, 5:e10439.

98. Editorial.: Two sides of the coin. Nature genetics 2012, 44:1287.

99. Kondrashov A: Genetics: The rate of human mutation. Nature 2012, 488:467–8.

100. Goriely A, McVean GAT, Röjmyr M, Ingemarsson B, Wilkie AOM: Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. Science (New York, N.Y.) 2003, 301:643–6.

101. ORPHANET [ORPHANET report series; 2012, Prevalence of rare diseases: Bibliographic data (http://www.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_alphabetical_list.pdf).]

102. Guo HH, Choe J, Loeb LA: Protein tolerance to random amino acid change. Proceedings of the National Academy of Sciences of the United States of America 2004, 101:9205–10.

Chapter 2

A *de novo* paradigm for mental retardation

Lisenka ELM Vissers [1]*, Joep de Ligt [1]*, Christian Gilissen [1], Irene Janssen [1], Marloes Steehouwer [1], Petra de Vries [1], Bart van Lier [1], Peer Arts [1], Nienke Wieskamp [1], Marisol del Rosario [1], Bregje WM van Bon [1], Alexander Hoischen [1], Bert BA de Vries [1], Han G Brunner [1#] and Joris A Veltman [1#]

1. Department of Human Genetics - 855, Nijmegen Centre for Molecular Life Sciences and Institute for Genetic and Metabolic Disorders, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands.

* These authors contributed equally to this work

# These authors jointly directed this work

Nature Genetics 2010; 42(12):1109-12

The term 'mental retardation' was used until 2010, when Rosa's Law was passed in the United States on October 5th, this law states that the term is replaced by 'intellectual disability'. This change of terminology has been almost universally accepted by the scientific community. The term 'mental retardation' is used in chapter 2 as we have left it in its original format, the publication was drafted before the law was passed and published as such.

The per-generation mutation rate in humans is high. *De novo* mutations may compensate for allele loss due to severely reduced fecundity in common neurodevelopmental and psychiatric diseases, explaining a major paradox in evolutionary genetic theory. Here we used a family based exome sequencing approach to test this *de novo* mutation hypothesis in ten individuals with unexplained mental retardation. We identified and validated unique non-synonymous *de novo* mutations in nine genes. Six of these, identified in six different individuals, are likely to be pathogenic based on gene function, evolutionary conservation and mutation impact. Our findings provide strong experimental support for a *de novo* paradigm for mental retardation. Together with *de novo* copy number variation, *de novo* point mutations of large effect could explain the majority of all mental retardation cases in the population.

Recent studies [1,2] have indicated that humans have an exceptionally high per-generation mutation rate of between 7.6×10^{-9} and 2.2×10^{-8}. An average newborn is calculated to have acquired 50 to 100 new mutations in his or her genome, resulting in approximately 0.86 new amino-acid–altering mutations [2]. Spontaneous germline mutations can have serious phenotypic consequences when they affect functionally relevant bases in the genome. In fact, their occurrence may explain why diseases with a severely reduced fecundity remain frequent in the human population, especially when the mutational target is large and comprised of many genes. This would explain a major paradox in the evolutionary genetic theory of mental disorders [3,4]. In agreement with this hypothesis, *de novo* copy number variations (CNVs) are a known cause of schizophrenia, autism and mental retardation [5,6]. Much less is known about the frequency and impact of *de novo* point mutations in these common diseases. Whole genome or exome sequencing now permits the study of these mutations and their role in disease in a systematic genome-wide manner. This approach has recently been used to identify causative genes in several rare syndromes [1,7–10]. In addition, targeted resequencing of the coding exons of the X chromosome revealed nine genes associated with X-linked forms of mental retardation [11], showing the strength of these analyses in common diseases. In this study, we used a family based whole-exome sequencing approach to test the *de novo* mutation hypothesis in an unselected cohort of individuals with mental retardation.

We sequenced the exomes of ten case-parent trios. All cases, eight males and two females, had moderate to severe mental retardation and a negative family history. Clinical evaluation did not lead to a syndromic or etiologic diagnosis (Supplementary Online Note). Prior cytogenetic analysis showed normal chromosomes, and array-based genomic profiling did not reveal *de novo* or other CNVs associated with mental

Table 1: *Overview of all variants detected per patient and impact of the prioritization steps for selecting candidate non-synonymous de novo mutations.*

Trio	1	2	3	4	5	6	7	8	9	10	average
High confidence variant calls	20,810	21,658	21,338	22,647	17,694	22,333	21,369	22,658	24,085	22,962	21,755
After exclusion of nongenic, intronic & synonymous variants	5,556	5,665	5,691	5,991	4,607	5,567	5,716	5,628	5,985	5,994	5,640
After exclusion known variants	165	159	157	155	120	136	120	149	96	171	143
After exclusion inherited variants	4	7	3	7	7	2	2	6	6	7	5

retardation. In addition, fragile X syndrome was excluded by *FMR1* repeat expansion analysis. On average, we obtained 3.1 Gb of mappable sequence data per individual after exome enrichment (37 Mb of genomic sequence targeting ~18,000 genes) and sequencing on one quarter of a SOLiD sequencing slide (Supplementary Methods and Supplementary Table 1). Color space reads were mapped to the reference genome. On average, 79.6% of the bases originated from the targeted exome, with 90% of the targeted exons covered at least ten times. The median exon coverage was 42-fold, indicating that the majority of variants present in each exome could be robustly detected using a custom bioinformatic analysis pipeline (Supplementary Figure 1). On average, we identified 21,755 genetic variants per individual with high confidence (**Table 1** and Supplementary Figure 2).

We developed an automated prioritization scheme to systematically identify all candidate dominant *de novo* mutations in each affected individual (**Figure 1**). We first excluded all, intronic and synonymous variants other than those occurring at canonical splice sites. This first step reduced the number of candidates to an average of 5,640 non-synonymous and canonical splice site variants per affected individual. We further reduced this number to 143 by excluding all known, likely benign, variants by comparison with data from dbSNP database v130 and our in-house variant database. Next, we used the exome data from each case's parents to exclude all remaining inherited variants. This resulted in an average of five (with a range of two to seven) candidate *de novo* non-synonymous mutations per affected individual (**Table 1**).

For all 51 candidate mutations (Supplementary Table 2), we performed Sanger sequencing to (i) validate the mutations observed in the probands and (ii) validate the absence of the mutations in the parental DNA. Thirty-eight candidates could not be validated in the proband (covered by a median of five variant reads in the exome sequencing experiment), but 13 candidates could be validated (covered by a median of 17 variant reads). Parental analysis validated the *de novo* occurrence for 9 of these 13 mutations, detected in seven different individuals (**Table 2** and Supplementary Figures 3 & 4). We did not identify these mutations in a total of 1,664 control chromosomes, nor did we see other likely pathogenic mutations identified in the affected genes in these control chromosomes, indicating that the population frequency of these types of *de novo* mutations in these genes will be lower than 0.22% (power=0.95, α=0.05). Eight of the *de novo* mutations were present in a heterozygous state on the autosomes and one was present in a hemizygous state on the X chromosome. All *de novo* mutations occurred in different genes, including two genes recently implicated in mental retardation (**Table 2**). In addition to using a dominant disease model, we also analyzed the data for recessive forms of mental retardation. In the affected male of trio 10, we identified a maternally inherited non-

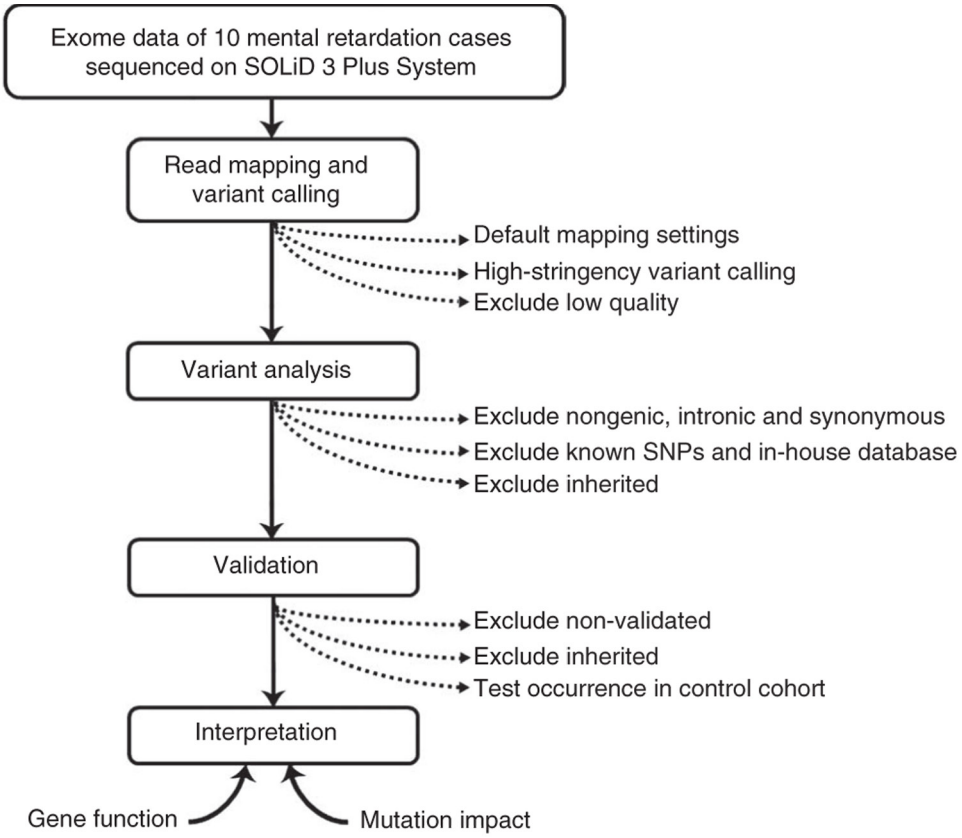


***Figure 1**, Experimental work flow for detecting and prioritizing sequence variants. For all ten mental retardation trios, prioritization of variants observed in the probands was based on selection for non-synonymous changes of high quality only and exclusion of all variants previously observed in healthy individuals, together with those variants that were inherited from an unaffected parent. Interpretation of de novo variants was based on gene function and the impact of the mutation.*

synonymous variant in *JARID1C* (**Table 2**), which is a well-described X-linked mental retardation gene [12]. Subsequent analysis of this variant in DNA obtained from the affected individual's grandparents indicated that the mutation had occurred *de novo* in the mother of this proband. No conclusive evidence for autosomal recessive inheritance, either homozygous or compound heterozygous, was obtained for the other affected individuals.

Next, we evaluated the function of each mutated gene in relation to the disorder (**Table 2**). Three genes do not seem to play a role in biological pathways linked to mental retardation. *BPIL3* is involved in the innate immune response [13], whereas *PGA5* is involved in protease activity in the stomach [14]. The function of *ZNF599*

BPIL3	3	M	NM_174897	c.887G>A	p.(Arg269His)	0.5	29	0.97	0.03	Innate immune respons
PGA5	4	F	NM_014224	c.1058T>C	p.(Val353Ala)	0.7	64	0.84	0.16	Precursor of pepsin
DEAF1	5	M	NM_021008	c.683T>G	p.(Ile228Ser)	4.9	142	0.01	0.99	Transcription factor; Regulator of 5-HT1A receptor in the brain; Mouse knockout causes neural tube defects
CIC	6	M	NM_015125	c.1474C>T	p.(Arg492Trp)	2.6	101	0.46	0.54	Granule cell development in central nervous system
SYNGAP1	8	F	NM_006772	c.998_999del	p.(Val333AlafsX)	3.3	-	-	-	Known autosomal dominant mental retardation gene
X-linked inherited muatations										
JARID1C	10	M	NM_001146702	c.1919G>A	p.(Cys640Tyr)	5.1	194	2.09×10^{-6}	1.00	Known X-linked mental retardation gene

 # Gender of proband, with M for male and F for female.
** Visual representation of probabilities are provided in Supplementary Figure 5.*
- Grantham scores for nonsense (in RAB39B) and frameshift mutations (in SYNGAP1) could not be calculated.

Table 2: *Overview of all de novo variants identified by exome sequencing in ten individuals with unexplained mental retardation.*

Gene	Trio	Sex#	NM number	cDNA level change	Protein level change	PhyloP score	Grantham score	Probability of being observed in dbSNP*	Probability of being observed in HMGD*	Gene function
					***De novo* mutations**					
DYNC1H1	1	M	NM_001376	c.11465A>C	p.(His3822Pro)	5.5	77	0.20	0.80	Retrograde axonal transporter; interacts with *PAFAH1B1* (causing Lissencephaly, a neurodevelopmental disorder)
ZNF599	1	M	NM_0010072488	c.532C>T	p.(Leu187Phe)	-1.5	22	1.00	2.65×10^{-4}	Unknown
RAB39B	2	M	NM_171998	c.557G>A	p.(Trp186X)	4.8	-	-	-	Known X-linked mental retardation gene
YY1	3	M	NM_003403	c.1138G>T	p.(Asp380Tyr)	6.9	160	2.27×10^{-6}	1.00	Ubiquitously expressed transcription factor; Mouse knock down results in growth retardation, neurulation defects and brain abnormalities; interacts with MeCP2, a known mental retardation gene

is currently unknown. For the six other genes affected by *de novo* mutations, functional evidence suggests a role in mental retardation. Two mutations occurred in genes (*RAB39B* and *SYNGAP1*) that, when disrupted, are known to cause mental retardation (**Table 2**) [15,16]. For the remaining four mutated genes, evidence for a causal link with mental retardation is provided by model organisms and protein-protein interaction studies. *DYNC1H1* encodes a cytoplasmic dynein that acts as a motor for intracellular retrograde axonal transport. Heterozygous *Dync1h1*+/− mutant mice exhibit sensory neuropathy [17], and studies in zebrafish have shown the importance of dync1h1 in correct nuclear positioning. Mislocalization of nuclei in the vertebrate central nervous system is likely to result in profound patterning defects and severely compromised function [18]. Notably, *DYNC1H1* interacts with *PAFAH1B1*, the gene associated with type I lissencephaly, which involves gross disorganization of the neurons within the cerebral cortex [19]. *YY1* encodes the ubiquitously expressed transcription factor yin-yang 1 and directs histone deacetylases and histone acetyltransferases, implicating chromatin remodeling as its main function. Complete ablation of *Yy1* in mice results in early embryonic lethality, whereas *Yy1* heterozygous mice display growth retardation, neurulation defects and brain abnormalities [20]. Recent studies show that YY1 interacts directly with MECP2; *MECP2* is mutated in Rett syndrome [21]. *DEAF1* encodes a transcription factor that regulates the 5-HT1A receptor in the human brain. Mutations in the Drosophila *DEAF1* ortholog result in early embryonic arrest, suggesting an essential role for the gene in early development [22]. Additional evidence is provided by *Deaf1*-deficient mice, which show neural tube defects including exencephaly [23]. Finally, *CIC* is a member of the HMG-box transcription factor superfamily, which is associated with neuronal and glial development of the nervous system. *CIC* is predominantly and transiently expressed in immature granule cells of the cerebellum, hippocampus and neocortex, suggesting a critical role in central nervous system development [24].

We next examined the evolutionary conservation of affected nucleotides (using the phyloP score), as well as the potential of the *de novo* mutations to affect the structure or function of the resulting proteins (using the Grantham score; **Table 2**). All *de novo* missense mutations and the inherited X-linked mutation were included in this analysis; no Grantham scores were available for the additional nonsense and frameshift mutations. Of note, *de novo* mutations in genes with a functional link to mental retardation showed a higher phyloP (mean, 4.7) and Grantham score (mean, 135) than mutations in genes without such a functional indication (mean phyloP score −0.5 and mean Grantham score 38). We also compared these scores to those for all non-synonymous variants in the dbSNP database as well as those in the Human Gene Mutation Database (HGMD). The distribution of phyloP scores

and Grantham scores differed markedly between dbSNP and the HGMD (**Online Methods** and Supplementary Figure 5). The four mutations in genes functionally linked to mental retardation all showed higher probability values for being observed in HGMD (mean 0.83) than for being observed in dbSNP (mean 0.17). The three mutations in genes without a functional link to mental retardation showed an average probability of 0.94 for being observed in dbSNP and an average probability of 0.06 for being observed in HGMD (**Table 2**). Additionally, the inherited *JARID1C* mutation showed a probability of 1.00 for being in HGMD versus 2.09×10^{-6} for being in dbSNP.

This analysis of the mutated nucleotides and their impact on gene function strongly supports pathogenicity for six of the nine *de novo* mutations. Importantly, these six mutations occurred in genes with a functional link to mental retardation, two of which are known mental retardation genes. In contrast, three *de novo* variants in genes without a functional link did not appear to significantly affect protein function. Moreover, we identified a maternally inherited mutation in a known X-linked mental retardation gene that arose *de novo* in the proband's mother. Although we have not provided individual functional tests to prove causality, these data collectively provide strong evidence for a major role of *de novo* mutations in mental retardation. The identification of recurrent mutations in these genes in unrelated cases would provide additional proof for disease causality, but this may require the evaluation of thousands of affected individuals. The identification of subtle CNVs encompassing (part of) these genes may also provide additional proof for disease causality, as was shown recently for mutations in X-linked mental retardation genes [25]. As of yet, no such CNVs have been reported, nor have we found such CNVs in our diagnostic cohort of ~4,500 individuals with mental retardation (data not shown).

The discovery of nine *de novo* non-synonymous mutations in this cohort of ten affected individuals is concordant with the recently estimated background mutation rate of 0.86 amino-acid–altering mutations per newborn in controls [2], but it will be important to compare this result to similar data from healthy control trios when available. Notably, after applying the same systematic filtering approach and Sanger sequencing, we could only validate a single *de novo* synonymous mutation, which occurred in *GRIN1* (c.351C>T, seen in trio 10). This base pair is not conserved through evolution (phyloP score=−3.2) and does not seem to alter splicing, suggesting that this mutation is an unlikely candidate for causing mental retardation.

Of note, the individual carrying this mutation also carries the *JARID1C* mutation. The observed ratio of non-synonymous to synonymous *de novo* mutations is far greater than would be expected for protein-coding genes under purifying selection and indicates that many of these mutations will result in a reproductive disadvantage. In contrast, the average non-synonymous to synonymous ratio reported in dbSNP

for the six genes with predicted pathogenic mutations is significantly lower than that of the three genes with mutations reflecting the background mutation rate (Fisher's Exact test, p=0.0016), which is to be expected for disease genes in the normal population.

In summary, our results suggest that *de novo* mutations are a major cause of unexplained mental retardation. These mutations can readily be identified using a family based exome sequencing approach and require only limited follow-up by Sanger sequencing. Our findings have implications for preventive and diagnostic strategies in mental retardation. Systematic genome-wide resequencing in parent-child trios may uncover further examples of this *de novo* paradigm for other human neurodevelopmental disorders.

URLs

1000 Genomes Project	http://www.1000genomes.org
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
HGMD	http://www.hgmd.cf.ac.uk/ac/index.php
R	http://www.r-project.org/

ACCESSION CODES

The genomic reference sequence for *DYNC1H1* can be found under the GenBank accession number NM_001376; for *ZNF599* under NM_001007248; for *RAB39B* under NM_171998; for *YY1* under NM_003403; for *BPIL3* under NM_174897; for *PGA5* under NM_014224; for *DEAF1* under NM_021008; for *CIC* under NM_015125; for *SYNGAP1* under NM_006772; for *JARID1C* under NM_001146702; and for *GRIN1* under NM_021569.2.

ACKNOWLEDGEMENTS

We thank R. de Reuver and J. Y. Hehir-Kwa for bioinformatics support in data analysis and personnel from the Sequencing Facility of our department for timely completion of Sanger sequencing of validation experiments. This work was funded in part by grants from The Netherlands Organization for Health Research and Development (ZonMW grants 916-86-016 to L.E.L.M.V., 917-66-36 and 911-08-025 to J.A.V. and 917-86-319 to B.B.A.d.V.), the EU-funded TECHGENE project (Health-F5-2009-223143 to J.d.L. and J.A.V.) and the AnEUploidy project (LSHG-CT-2006-37627 to A.H., B.W.M.v.B., H.G.B., B.B.A.d.V. and J.A.V.).

REFERENCES

1. Roach, J.C. et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328, 636–639 (2010).

2. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. USA 107, 961–968 (2010).
3. Keller, M.C. & Miller, G. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? Behav. Brain Sci. 29, 385–404 (2006).
4. Uher, R. The role of genetic variation in the causation of mental illness: an evolution-informed framework. Mol. Psychiatry 14, 1072–1082 (2009).
5. Cook, E.H. Jr. & Scherer, S.W. Copy-number variations associated with neuropsychiatric conditions. Nature 455, 919–923 (2008).
6. de Vries, B.B. et al. Diagnostic genome profiling in mental retardation. Am. J. Hum. Genet. 77, 606–616 (2005).
7. Ng, S.B. et al. Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. 42, 30–35 (2010).
8. Lupski, J.R. et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N. Engl. J. Med. 362, 1181–1191 (2010).
9. Hoischen, A. et al. *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat. Genet. 42, 483–485 (2010).
10. Sobreira, N.L. et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet. 6, e1000991 (2010).
11. Tarpey, P.S. et al. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat. Genet. 41, 535–543 (2010).
12. Jensen, L.R. et al. Mutations in the JARID1C gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-linked mental retardation. Am. J. Hum. Genet. 76, 227–236 (2005).
13. Mulero, J.J. et al. Three new human members of the lipid transfer/lipopolysaccharide binding protein family (LT/LBP). Immunogenetics 54, 293–300 (2002).
14. Taggart, R.T. et al. Relationships between the human pepsinogen DNA and protein polymorphisms. Am. J. Hum. Genet. 38, 848–854 (1986).
15. Giannandrea, M. et al. Mutations in the small GTPase gene RAB39B are responsible for X-linked mental retardation associated with autism, epilepsy, and macrocephaly. Am. J. Hum. Genet. 86, 185–195 (2010).
16. Hamdan, F.F. et al. Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. N. Engl. J. Med. 360, 599–605 (2009).
17. Chen, X.J. et al. Proprioceptive sensory neuropathy in mice with a mutation in the cytoplasmic Dynein heavy chain 1 gene. J. Neurosci. 27, 14515–14524 (2007).
18. Tsujikawa, M., Omori, Y., Biyanwila, J. & Malicki, J. Mechanism of positioning the cell nucleus in vertebrate photoreceptors. Proc. Natl. Acad. Sci. USA 104, 14819–14824 (2007).
19. Tai, C.Y., Dujardin, D.L., Faulkner, N.E. & Vallee, R.B. Role of dynein, dynactin, and CLIP-170 interactions in LIS1 kinetochore function. J. Cell Biol. 156, 959–968 (2002).
20. He, Y. & Casaccia-Bonnefil, P. The Yin and Yang of YY1 in the nervous system. J. Neurochem. 106, 1493–1502 (2008).

21. Forlani, G. et al. The MeCP2/YY1 interaction regulates ANT1 expression at 4q35: novel hints for Rett syndrome pathogenesis. Hum. Mol. Genet. 19, 3114–3123 (2010).
22. Veraksa, A., Kennison, J. & McGinnis, W. DEAF-1 function is essential for the early embryonic development of Drosophila. Genesis 33, 67–76 (2002).
23. Hahm, K. et al. Defective neural tube closure and anteroposterior patterning in mice lacking the LIM protein LMO4 or its interacting partner Deaf-1. Mol. Cell. Biol. 24, 2074–2082 (2004).
24. Lee, C.J. et al. CIC, a member of a novel subfamily of the HMG-box superfamily, is transiently expressed in developing granule neurons. Brain Res. Mol. Brain Res. 106, 151–156 (2002).

ONLINE METHODS

SUBJECTS

Ten individuals with unexplained moderate to severe mental retardation (with normal karyotypes and genomic profiles obtained using 250K SNP arrays) were selected for exome sequencing (Supplementary Note). Family history for mental retardation was negative for all cases. Nongenic causes for mental retardation, including pre-, peri- and post-natal infection and perinatal injury, were excluded. DNA was obtained from peripheral blood from the ten probands as well as from their unaffected parents. DNA isolation was performed using QIAamp DNA Mini Kit (QIAGEN), according to the instructions of the manufacturer. This study was approved by the Medical Ethics Committee of the Radboud University Nijmegen Medical Centre, and all participants signed written informed consent.

LIBRARY GENERATION

Exome enrichment required 3 µg of genomic DNA, and an AB SOLiD Optimized SureSelect Human Exome Kit (Agilent) was used for enrichment, containing the exonic sequences of ~18,000 genes and covering a total of ~37 Mb of genomic sequence, as specified by the company. We followed the manufacturer's instructions (version 1.5) for enrichment with a minor modification, which was the reduction of the number of post-hybridization ligation-mediated PCR cycles from 12 cycles to 9 cycles.

SOLID SEQUENCING

The enriched exome libraries were subsequently used for emulsion PCRs, following the manufacturer's instructions (Life Technologies), based on a library concentration of 1 picomolar (pM) (version March 2010). For each sample, one-quarter of a sequencing slide (Life Technologies) was used on a SOLiD 3 Plus System.

MAPPING OF VARIANTS

Color space reads were mapped to the hg18 reference genome with the SOLiD bioscope software v1.2, which utilizes an iterative mapping approach. Single-nucleotide variants were subsequently called by the diBayes algorithm26 using high stringency settings, requiring calls on each strand. Small insertions and deletions were detected using the SOLiD Small Indel Tool. We assumed a binomial distribution with a probability of 0.5 of sequencing the variant allele at a heterozygous position. Under this assumption, at least ten reads are required to obtain a 99% probability that at least two reads contain the variant allele. Variants and indels were selected using strict quality control settings, which included the presence of at least four unique variant reads (that is, having different start sites), as well as the variant being present in at least 15% of all reads. All called variants and indels were combined and annotated using a custom analysis pipeline (resulting in HCDiff files for each individual).

CUSTOM BIOINFORMATIC ANALYSIS PIPELINE

All variants reported in the HCDiff files were filtered to ensure an optimal prioritization process. For this, we first excluded all nongenic, intronic (other than canonical splice sites) and synonymous variants, reducing the number of variants to an average of 5,640 per individual. Second, all known variants were excluded by comparison with data from dbSNP v130 as well as from our in-house variant database. At the time of this study, this in-house database contained variants from (i) 78 in-house performed 'exomes', contributing 515,480 variants, and (ii) the 1000 Genomes Project (see URLs) and published data from various other studies [27–29], contributing 3,059,835 variants, thereby bringing the number of variants in the in-house database to 3,525,278. Of note, if the variant observed in the proband occurred at a genomic position known in dbSNP v130, but the change present was different in the two (for example, A/C in dbSNP but A/T in the proband), the variant was not excluded from analysis. The filtering step using this data further reduced the average number of variants to 143 per proband.

Next, for a dominant model of disease, we used the exome data from accompanying parents to exclude all inherited variants. This step further reduced the number of potential *de novo* variants to an average of 33 per proband. As not all variants identified in the exomes of the probands may have been sequenced at sufficient coverage in the parental samples, we checked all remaining variants in the exome data from the accompanying parents. In brief, even if only a single read showed the variant allele in one of the parental exome samples, the variant was excluded for validation in the proband. Simultaneously, we checked all remaining potential *de novo* indels for annotation differences in each child-parent trio and excluded

those that were found to be identical variants in both parent and child. After this final check, an average of five potential *de novo* variants per proband remained for further validation.

To evaluate the presence of recessive mutations, variant filtering was essentially performed as described above, with the main difference being that uniquely inherited parental variants were not excluded here. The remaining variants were evaluated for the presence of compound heterozygous variants, as well as variants that were present in >80% of all reads. Subsequently, parental exome data were used for segregation analysis of the variants identified.

DBSNP AND HGMD

To explore the pathogenicity of our *de novo* variants, the genomic evolutionary conservation score (phyloP) and the amino-acid change (Grantham) were compared to those scores present in dbSNP (build 130) and the HGMD (see URLs). All non-synonymous changes reported in dbSNP and HGMD were retrieved, and overlap between databases was removed from both datasets. In addition, non-synonymous variants in dbSNP with an OMIM disease entry, suggestive for a Mendelian phenotype, were omitted from the dbSNP dataset.

Next, quadratic discriminant analysis [30] was performed on these two datasets to determine the significance of the phyloP and Grantham scores as discriminating factors. Statistical tests were performed using the R statistics package (see URLs). The assumption of normality in the data required for the model was determined using Lilliefors (Kolmogorov-Smirnov) normality testing [31]: PhyloP $D=0.0626$, $P<2.2 \times 10^{-16}$; Grantham $D=0.0828$, $P<2.2 \times 10^{-16}$; PhyloP × Grantham $D=0.1395$, $P<2.2 \times 10^{-16}$. D represents the maximum absolute difference between the empirical and hypothetical cumulative distribution function.

The combination of both scores together yielded the highest power to discriminate the two datasets, and as such, the combined value was used to calculate probabilities for our *de novo* variants to be observed in either database.

VALIDATION EXPERIMENTS

Validation and *de novo* testing for candidate *de novo* mutations was performed using standard Sanger sequencing approaches. Primers were designed to surround the candidate mutation, and PCR reactions were performed using RedTaq Readymix PCR reaction mix (Sigma-Aldrich). Primer sequences and PCR conditions are available upon request. For all *de novo* mutations identified, an additional control cohort of 75 ethnically matched controls was tested for the presence of the same mutation by Sanger sequencing. Together with the results from 679 control individuals from the 1000 Genomes Project as well as the 78 ‘exomes’ present in our in-house

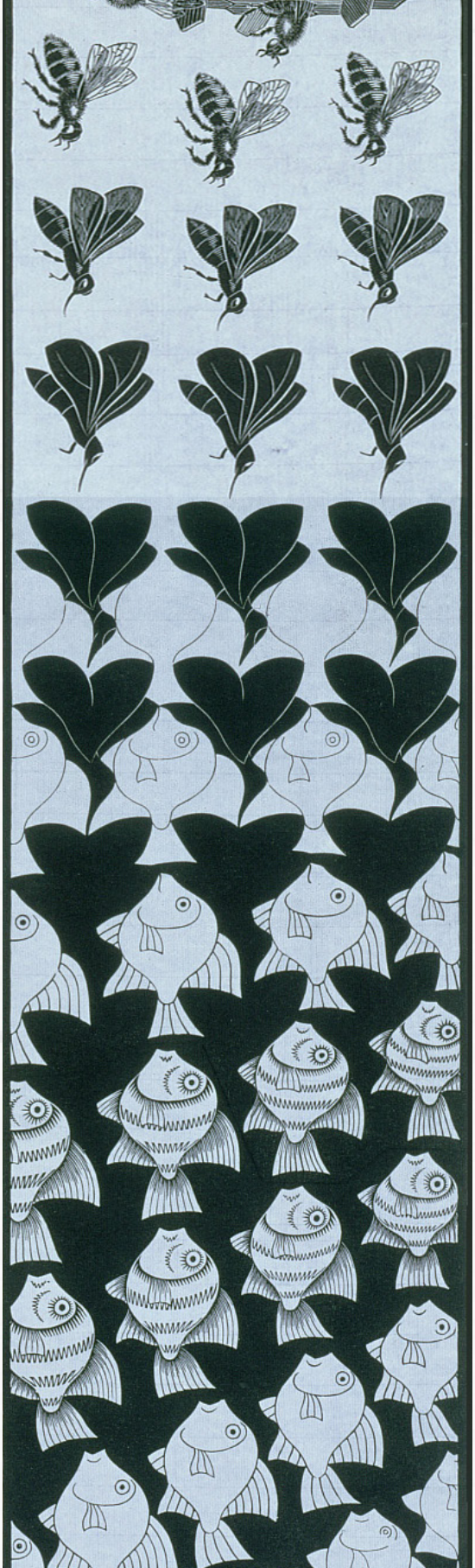
database, the control cohort for the *de novo* mutations encompassed 1,664 control chromosomes.

ADDITIONAL REFERENCES

25. Whibley, A.C. et al. Fine-scale survey of X chromosome copy number variants and indels underlying intellectual disability. Am. J. Hum. Genet. 87, 173–188 (2010).
26. Marth, G.T. et al. A general approach to single-nucleotide polymorphism discovery. Nat. Genet. 23, 452–456 (1999).
27. Ng, S.B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276 (2009).
28. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. Nat. Biotechnol. 27, 847–852 (2009).
29. Wang, J. et al. The diploid genome sequence of an Asian individual. Nature 456, 60–65 (2008).
30. Venables, W.N. & Ripley, B.D. Modern Applied Statistics with R (Springer, 4th edn., New York, New York, USA, 2002).
31. Lilliefors, H. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. 62, 399–402 (1967).

SUPPLEMENTARY TABLES AND FIGURES

Supplementary data is freely available online; http://www.nature.com/ng/journal/v42/n12/abs/ng.712.html#supplementary-information



Chapter 3

Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability

Joep de Ligt[1]*, Marjolein H. Willemsen[1]*, Bregje W.M. van Bon[1]*, Tjitske Kleefstra[1]*, Helger G. Yntema[1], Thessa Kroes[1], Anneke T. Vulto-van Silfhout[1], David A. Koolen[1], Petra de Vries[1], Christian Gilissen[1], Marisol del Rosario[1], Alexander Hoischen[1], Hans Scheffer[1], Bert B.A. de Vries[1], Han G. Brunner[1], Joris A. Veltman[1#] and Lisenka E.L.M. Vissers[1#]

1. Department of Human Genetics, Nijmegen Center for Molecular Life Sciences, Institute for Genetic and Metabolic Disease, Radboud University Nijmegen Medical Center, PO Box 9101, 6500 HB Nijmegen, The Netherlands.

* these authors contributed equally to this article

# these authors contributed equally to this article

New England Journal of Medicine 2012 Nov 15;367(20)

ABSTRACT

BACKGROUND

The causes of intellectual disability remain largely unknown because of extensive clinical and genetic heterogeneity.

METHODS

We evaluated patients with intellectual disability to exclude known causes of the disorder. We then sequenced the coding regions of more than 21,000 genes obtained from 100 patients with an IQ below 50 and their unaffected parents. A data-analysis procedure was developed to identify and classify *de novo*, autosomal recessive, and X-linked mutations. In addition, we used high-throughput resequencing to confirm new candidate genes in 765 persons with intellectual disability (a confirmation series). All mutations were evaluated by molecular geneticists and clinicians in the context of the patients' clinical presentation.

RESULTS

We identified 79 *de novo* mutations in 53 of 100 patients. A total of 10 *de novo* mutations and 3 X-linked (maternally inherited) mutations that had been previously predicted to compromise the function of known intellectual-disability genes were found in 13 patients. Potentially causative *de novo* mutations in novel candidate genes were detected in 22 patients. Additional *de novo* mutations in 3 of these candidate genes were identified in patients with similar phenotypes in the confirmation series, providing support for mutations in these genes as the cause of intellectual disability. We detected no causative autosomal recessive inherited mutations in the discovery series. Thus, the total diagnostic yield was 16%, mostly involving *de novo* mutations.

CONCLUSIONS

De novo mutations represent an important cause of intellectual disability; exome sequencing was used as an effective diagnostic strategy for their detection. (Funded by the European Union and others.)

INTRODUCTION

Severe intellectual disability, which is also referred to as cognitive impairment or mental retardation, affects approximately 0.5% of the population in Western countries [1,2] and represents an important health burden. A clinical diagnosis of severe intellectual disability is generally based on an IQ of less than 50 and substantial limitations in activities of daily living. In early childhood, the diagnosis is

based on substantial developmental delays, including motor, cognitive, and speech delays. Children with different nonsyndromic forms of intellectual disability are clinically indistinguishable.

Intellectual disability can be caused by nongenetic factors, such as infections and perinatal asphyxia. In developed countries, most severe forms of intellectual disability are thought to have a genetic cause [2], but the cause remains elusive in 55 to 60% of patients [3,4]. An understanding of the genetic cause of intellectual disability can benefit patients and their families, because a diagnosis may provide information on the prognosis, precludes further unnecessary invasive testing, and may lead to appropriate therapy. Moreover, a diagnosis often facilitates access to appropriate medical and supportive care [5-8]. Family members may benefit from knowledge of the risk of recurrence, reproductive counseling, and possible prenatal diagnosis.

We [9] and others [10] have reported evidence supporting the hypothesis that rare *de novo* point mutations can be a major cause of severe intellectual disability. Recent studies have indicated that there are more *de novo* mutations in persons with intellectual disability than in healthy controls, highlighting the clinical importance of these mutations [11-15]. That intellectual disability is often sporadic, without obvious environmental or familial factors, provides additional support for the hypothesis that a large proportion of cases of intellectual disability are caused by *de novo* mutations. It has been estimated that mutations in more than 1000 different genes may cause intellectual disability [16]. Because of this large aggregate target, rare *de novo* mutations in these genes may collectively compensate for the very low rate of reproduction among patients with intellectual disability, keeping the incidence of the disorder in the general population stable [15].

In the absence of diagnostic clues from the clinical phenotype, unbiased genome-wide approaches are required to detect genetic mutations causing intellectual disability [9,17,18]. We have therefore evaluated the role of *de novo* as well as X-linked and autosomal recessive inherited mutations in a series of 100 patients with unexplained intellectual disability defined as an IQ of <50), using a family-based exome-sequencing approach in a clinical diagnostic setting. Previous extensive clinical and genetic evaluation of these patients had not led to an etiologic diagnosis. Thus, this series of patients represents the end point of current diagnostic strategies, with all conventional genetic resources exhausted, which is the typical scenario for patients with severe intellectual disability [19].

***Table 1**, Clinical characteristics of 100 patients with Intellectual Disability of unknown cause.*

Characteristic	No. of Patients	Characteristic	No. of Patients
IQ		***Short stature***	
<30	62	Yes	24
30 to 50	38	No	76
Gender		***Microcephaly***	
Male	47	Yes	30
Female	53	No	70
Age group		***Macrocephaly***	
<10 yr 37	37	Yes	4
10–20 yr	41	No	96
>20 yr	22	***Epilepsy***	
No. of siblings		Yes	52
0	12	No	48
1	47	***Abnormality on brain imaging***	
2	36	Yes	30
3	1	No	40
4	2	Not assessed	30
Unknown	2	***Cardiac malformation***	
No. of major congenital anomalies		Yes	2
0	62	No	98
1	31	***Abnormality of the urogenitary system***	
2	7	Yes	13
3	0	No	87
No	48		

METHODS

PATIENTS

We enrolled 100 patients (53 females and 47 males) with unexplained severe intellectual disability and their unaffected parents (trios). This series is broadly representative of patients with severe intellectual disability who are referred to our tertiary care clinic (see Table S1). All patients were evaluated by a clinical geneticist. Detailed clinical phenotypes of the 100 patients are provided in the section on the clinical descriptions of patients in the Online Supplementary Appendix and are summarized in **Table 1**. Before enrollment, the patients had undergone an extensive diagnostic workup, including genomic profiling (performed with the use of the 250K Affymetrix SNP array) and targeted gene tests, with metabolic screening whenever

indicated, but these evaluations had not led to an etiologic diagnosis. The study was approved by the ethics committee at the Radboud University Nijmegen Medical Center. The parents of all patients in the study provided written informed consent.

DETECTION OF MUTATIONS

Genomic DNA was isolated from blood with the use of a QIAamp DNA Mini Kit (Qiagen). Exomes were enriched with the use of a SOLiD-Optimized SureSelect Human Exome Kit (Agilent version 2, 50 Mb), followed by SOLiD 4 System sequencing (Life Technologies). After sequencing the exomes of each trio, we selected candidate *de novo* mutations by excluding variants inherited from either parent and selected candidate recessive and X-linked mutations by segregation analysis (Figure S1 in the Supplementary Appendix). Candidate *de novo* mutations were validated by conventional Sanger sequencing methods in DNA samples obtained from the patients and their parents (see the section on validation of *de novo* mutations in the Supplementary Appendix).

TESTING OF CANDIDATE GENES

We reanalyzed all candidate genes that were identified in this study for the presence of possible *de novo* mutations in previously generated exome data obtained from 10 patients with severe intellectual disability [9]. In addition, we resequenced five candidate genes associated with intellectual disability (*DYNC1H1, KIF5C, ASH1L, GATAD2B,* and *CTNNB1*) using array-based enrichment on pooled DNA samples from a second series of 765 patients with intellectual disability. These samples were selected from our in-house collection of 5,621 samples from patients with undiagnosed intellectual disability (see the section on patient selection in the Supplementary Appendix). The parents of these patients had previously provided written informed consent. All patients had been evaluated by a clinical geneticist to rule out known causes of intellectual disability, and genomic array analysis had not revealed causal copy-number variants. Detected variants were annotated and prioritized according to their presumed relevance to disease. Variants fulfilling prioritization criteria were validated by means of conventional Sanger sequencing (see the section on recurrence screening in the Supplementary Appendix).

INTERPRETATION OF CONFIRMED MUTATIONS

We classified the mutations on the basis of the existing guidelines for evaluation of the pathogenicity of variants [20,21] (**Figure 1**, and the section on clinical interpretation of mutations in the Supplementary Appendix). These guidelines call for the evaluation of seven factors: the function of the affected gene, the effect of the mutation on the codon (i.e., stop, frameshift, or missense mutation), in silico

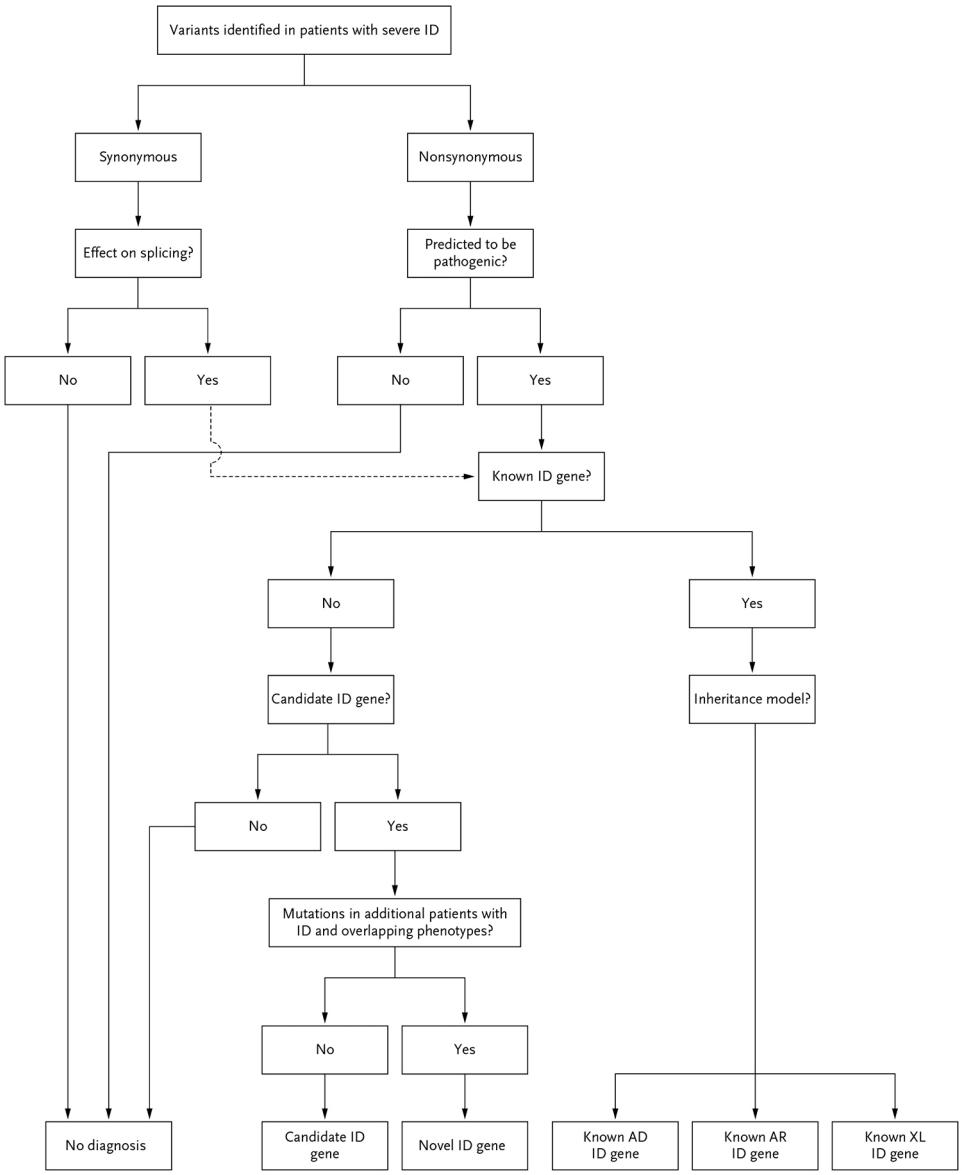


Figure 1. *Classification of Variants Detected in Patients with Severe Intellectual Disability (ID). Variants in known autosomal recessive (AR) genes were considered to be diagnostically relevant only if biallelic, predicted pathogenic variants were detected. AD denotes autosomal dominant, and XL X-linked.*

prediction of the functional effect at the amino acid level, evolutionary conservation, brain-expression patterns, analysis of gene ontology (GO) terms, and the use of animal models, if available (see the section on mutational effect and functional relevance in the Supplementary Appendix).

Mutations in genes with a known association with intellectual disability were considered to be a cause of intellectual disability when the mutations were predicted to be pathogenic by the majority of three prespecified in silico analyses (see the section on mutational effect in the Supplementary Appendix) and when they occurred in persons with phenotypes similar to those described in other persons with mutations in these genes.

Mutations were considered to affect candidate genes that had not previously been implicated in intellectual disability when the mutations were predicted to be pathogenic by the majority of three prespecified in silico analyses, showed a link to brain or embryonic development in a review of the literature, and met at least two of the following criteria: evolutionary conservation, brain-expression pattern, positive results on GO term analysis, or implication on the basis of animal models (see the section on functional relevance in the Supplementary Appendix). If multiple patients were found to have a *de novo* mutation in such a candidate gene and their phenotypes showed striking overlap, the candidate gene was redefined as a novel intellectual disability gene, and the mutations were reported as a cause of intellectual disability. All mutations in other candidate genes were reported as a possible cause of intellectual disability. For patients without causal *de novo*, X-linked, or biallelic inherited mutations, the diagnostic report stated that the genetic cause of intellectual disability was not identified.

RESULTS

EXOME SEQUENCING

The power of family-based exome sequencing to provide a genetic diagnosis was evaluated for 100 patients with unexplained intellectual disability. The median sequence coverage was 64, with an average of 87% of all targeted exons covered by at least 10 sequence reads (Table S1 & Figure S2 in the Supplementary Appendix). We detected an average of 24,324 genetic variants per patient (Table S2 in the Supplementary Appendix). An automated prioritization scheme was applied to systematically identify candidate *de novo* mutations (Figure S1 in the Supplementary Appendix), resulting in a total of 690 candidate *de novo* mutations (average number per patient, 7; range, 2 to 20) (Table S2 in the Supplementary Appendix).

An inherent challenge in family-based exome sequencing is the difficulty in distinguishing between a true *de novo* mutation and a sequencing error, since

both appear to be a new allele in the patient. Therefore, we tested the veracity of all candidate *de novo* mutations using Sanger sequencing as an independent method and confirmed the presence of 79 *de novo* mutations in 53 patients (range per patient, 1 to 4) (Tables S3 and S4 & Figure S3 and S4 in the Supplementary Appendix).

DESCRIPTION OF *DE NOVO* MUTATIONS

Of the 79 *de novo* mutations (affecting 77 genes), 16 were synonymous (**Figure 1**). Of these mutations, none were predicted to alter splicing. Therefore, these mutations were classified as not causative for intellectual disability. The remaining 63 changes were nonsynonymous and included 15 mutations that were predicted to be severely disruptive (4 nonsense mutations, 2 canonical splice-site mutations, and 9 insertions or deletions) and 48 missense mutations (Table S3 in the Supplementary Appendix). On the basis of random mutation modeling, 22 our observation of 4 nonsense mutations (6.1%) exceeded expectations (3.3%) and thus supports an elevated level of such mutations in patients with intellectual disability (Table S5 in the Supplementary Appendix). Prediction of all 48 missense mutations with the use of Polymorphism Phenotyping, version 2 (PolyPhen2), showed a significant increase in mutations that were probably damaging (p=0.03) (Table S5 in the Supplementary Appendix). This finding suggested that a large portion of these missense mutations might have phenotypic consequences.

We detected 12 *de novo* mutations in known intellectual-disability genes: 6 severely disruptive mutations and 6 missense mutations (**Table 2** & Table S3 in the Supplementary Appendix). Three *de novo* mutations were detected in genes known to cause a recessive form of intellectual disability when mutated. These mutations would be considered causal for intellectual disability only if a second, inherited mutation that was predicted to be pathogenic was identified. We identified no such mutation in *ARFGEF2* or *TUSC3*. However, we did detect a rare paternally inherited, predicted pathogenic variant (c.6160G>A; p.(Asp2054Asn)) in *LRP2*. The results of analyses performed to determine whether the *de novo* event occurred on the maternal haplotype were inconclusive. Recessive *LRP2* mutations cause the Donnai–Barrow syndrome, and clinical reevaluation of Patient 81 confirmed this diagnosis (Table S6 and the Clinical Description of Patients section in the Supplementary Appendix).

We analyzed the remaining 51 *de novo* mutations and identified 25 mutations in 24 candidate genes (Table S3 in and the Supplementary Appendix). A patient with a *de novo DYNC1H1* mutation and intellectual disability has been described previously [9]. A comparison between that patient and Patient 54 in our study showed that they both had severe intellectual disability and a variable presentation of a neuronal

migration defect [23] (Figure S5 in the Supplementary Appendix).

ADDITIONAL PATIENTS WITH INTELLECTUAL DISABILITY

We reanalyzed previously generated exome data for 10 patients with undiagnosed severe intellectual disability [9] and resequenced five candidate genes associated with intellectual disability (*DYNC1H1*, *GATAD2B*, *ASH1L*, *KIF5C*, and *CTNNB1*) in a series of 765 persons with intellectual disability in order to identify additional *de novo* mutations (Tables S7 and S8 in the Supplementary Appendix).

In this confirmation series, we identified a second *de novo* mutation in the transcriptional repressor *GATAD2B*. The two *de novo* mutations that were observed in this gene, a nonsense p.(Gln470*) and a frameshift p.(Asn195Lysfs*30) mutation, both resulting in a stop codon (indicated by the star symbol), were predicted to be severely disruptive and to result in loss of function (Figure S6 in the Supplementary Appendix). Both patients with these mutations presented with severe cognitive and motor delays and limited speech, and the two patients had similar facial features. One additional severely disruptive *de novo* mutation was detected in *CTNNB1* (Figure S7 in the Supplementary Appendix). This mutation (p.(Arg515*)) and the *de novo* mutation detected on exome sequencing (p.(Ser425Thrfs*11)) were predicted to result in loss of function. A third patient carried a p.(Gln309*) mutation in *CTNNB1*. This mutation was not present in maternal DNA, and paternal DNA was unavailable. All three patients presented with severe intellectual disability, absent or very limited speech, microcephaly, and spasticity with a severely impaired ability to walk.

Patients 4 and 15 had *de novo* missense mutations in *TRIO*: p.(Asp1368Val) and p.(Thr2563Met), respectively. *TRIO* encodes a protein that acts in several signaling pathways that control cell proliferation [24]. These patients were not similar in any clinical respect other than intellectual disability (see the section on clinical descriptions in the Supplementary Appendix). Both patients also carried a mutation in a known intellectual-disability gene: *PDHA1* in Patient 4 and *TCF4* in Patient 15. Their phenotypes overlapped markedly with those of other patients with mutations in *PDHA1* and *TCF4* (Table S6 in the Supplementary Appendix), indicating that these mutations are the most likely cause of intellectual disability, although the mutations in *TRIO* may also play a part.

INHERITED MUTATIONS IN AUTOSOMAL RECESSIVE AND X-LINKED GENES

We detected 14 X-linked inherited mutations in 12 male patients (Table S9 in the Supplementary Appendix). Three of these mutations were located in known X-linked intellectual-disability genes (one in *PDHA1* and two in *ARHGEF9*). These mutations were predicted to be pathogenic, and the phenotypes that were observed in the patients were consistent with previous reports of affected persons carrying

mutations in these genes (Table S6 in the Supplementary Appendix). In 10 male patients, we also detected 11 X-linked inherited mutations in 11 genes that had not previously been associated with intellectual disability. Of these genes, *TRPC5* was classified as possibly causal. The analysis for autosomal recessive causes of intellectual disability revealed biallelic inherited mutations in 9 genes, including 2 genes (*PCNT* and *VPS13B*) that had previously been associated with an autosomal recessive form of intellectual disability. None of these mutations had been classified as a possible cause of intellectual disability (Table S9 in the Supplementary Appendix).

***Table 2**, Genes Affected by De Novo Mutations Associated with Intellectual Disability.*

Type of mutation	Known genes	Novel genes*	Candidate genes
Missense	*ARFGEF2 †, GRIN2A ‡, GRIN2B, TCF4, TUSC3 †*	*DYNC1H1*	*ASH1L, CAMKIIG, COL4A3BP, EEF1A2, GRIA1, KIF5C, LRP1, MIB1, PHACTR1, PPP2R5D, PROX2, PSMA7, RAPGEF1, TANC2, TNPO2, TRIO ‡*
Nonsense	*SCN2A*	*GATAD2B*	*PHIP, WAC*
Frameshift	*LRP2 §, PDHA1 SLC6A8, TUBA1A*	*CTNNB1*	*MTF1, ZMYM6*
Splice site	*SYNGAP1*		*MYT1L*

** Genes were defined as novel if there were additional de novo mutations in patients with phenotypic overlap. Details on de novo mutations are provided in Table S3 in the Supplementary Appendix.*

† This autosomal recessive gene was identified as mutated in a patient in whom no second mutation was detected.

‡ De novo mutations in this gene were found in two independent patients.

§ This autosomal recessive gene was found in a patient in whom a second rare, inherited mutation was detected.

FAMILY-BASED EXOME SEQUENCING

Conclusive genetic diagnoses were obtained for 10 patients with *de novo* mutations in known intellectual-disability genes and for 3 male patients with severely disruptive, maternally inherited mutations in known X-linked intellectual disability genes (Tables S3 and S9 in the Supplementary Appendix). The phenotypes of these patients fit well with previously reported phenotypes caused by mutations in these genes (Table S6 and the section on clinical descriptions in the Supplementary Appendix). No diagnostically relevant, inherited autosomal recessive mutations were identified. Thus, a diagnostic yield of 13% was obtained from mutations in

known intellectual-disability genes (Table S10 in the Supplementary Appendix).
Our study identified 24 novel candidate genes affected by *de novo* mutations. A pathogenic role for 3 of these genes was substantiated by the identification of additional patients with intellectual disability and severely disruptive mutations. In each case, there was striking phenotypic overlap observed among the patients with mutations in the same gene. We therefore conclude that *DYNC1H1*, *GATAD2B*, and *CTNNB1* are novel intellectual disability genes, which raises the diagnostic yield of exome sequencing to 16% (**Table 2** and **3**).

Table 3. *Diagnostic Yield of Exome Sequencing in the Patients.*

Positive Diagnosis	No. of Patients
All mutations	16
***De novo* mutations**	**13**
Autosomal dominant	10*
X-linked	2
Autosomal recessive	1†
Inherited mutations	**3**
X-linked	3
Autosomal recessive	0

** Seven patients had mutations in autosomal dominant genes that had previously been associated with intellectual disability, and three patients had mutations in novel autosomal dominant genes.*

† This patient had one de novo mutation and a second inherited, predicted pathogenic mutation.

DISCUSSION

Mutations in more than 400 genes have been linked to intellectual disability, but most of these mutations have a very low prevalence and their phenotypes are often indistinguishable. This argues for an unbiased diagnostic approach, especially since these 400 genes may represent less than half of all intellectual-disability genes. We implemented family-based diagnostic exome sequencing for patients with severe, unexplained intellectual disability. Exome sequencing is a procedure that is highly amenable to automation. Variants with potential clinical consequences can easily be validated with the use of Sanger sequencing as an independent method. We did not identify any major hurdles in the laboratory workflow in this study, which allowed for smooth integration of this process into diagnostics. *De novo* mutations

were present in 53% of the patients and provided a conclusive genetic diagnosis in at least 13%, with an additional 3% of X-linked inherited mutations in known intellectual-disability genes. This diagnostic yield is similar to that of current chromosomal analyses based on genomic arrays, and the two approaches are complementary [4,25-27]. We expect that the diagnostic rate of exome sequencing will increase with the identification of additional patients who have mutations in the novel candidate genes reported here.

The identification of causal mutations in known intellectual-disability genes in 16 of 100 patients provides clinically useful information for clinicians and for patients and their families, since much is known about the prognoses associated with these mutations. The identification of the underlying genetic cause may also lead to specific treatment options or dietary advice. As an example, a ketogenic diet was recommended for our patients with a mutation in *PDHA1* [28]. In addition, a specific anti-epileptic treatment, with the avoidance of sodium-channel blockers, was suggested for our patient with a *de novo SCN2A* mutation, since this therapy leads to better seizure control and improvement in cognitive functioning and quality of life in patients with *SCN1A* mutations [29].

Our studies suggest that several of the new candidate genes that we identified may be confirmed as having recurrent mutations in patients with intellectual disability. We already identified additional *de novo* mutations in three of five genes (*DYNC1H1*, *GATAD2B*, and *CTNNB1*) that were sequenced in a second set of affected persons, and detailed clinical analysis of these patients provided definitive evidence that these three genes cause intellectual disability when mutated. The identification of recurrently mutated genes in combination with detailed clinical phenotyping may reveal novel intellectual-disability genes and identify clinical subtypes of intellectual disability that may require specific approaches to clinical management. We observed evidence of autosomal recessive disease in only one affected patient, who carried a *de novo* mutation and a rare inherited mutation in *LRP2*. The apparent absence of pathogenic mutations in autosomal recessive intellectual-disability genes in our series suggests that this form of intellectual disability is rare in patients with isolated intellectual disability from nonconsanguineous parents. An analysis of referrals for intellectual disability to our tertiary care center showed that approximately 90% of patients have sporadic intellectual disability and nonconsanguineous parents (see the Supplementary Appendix). X-linked forms of intellectual disability were diagnosed in 5 of 100 patients, and in 2 of these 5 patients, the mutation occurred *de novo*. Mutations outside the coding regions, as well as mosaic, digenic, or oligogenic causes of intellectual disability, remain to be defined.

Unbiased diagnostic approaches such as exome sequencing may also reveal clinically relevant mutations that are not related to the disease under investigation.

An independent expert panel determined the clinical relevance of such incidental findings. Before study enrollment, all families were counseled about this possibility and consented to being informed if the findings were deemed to be relevant by this panel. No families objected to being informed about incidental findings. In this study, we encountered one incidental finding, a *de novo* c.517C>T (p.(Tyr173His)) change in *RB1*. Mutations in this gene are associated with retinoblastoma (Online Mendelian Inheritance in Man [OMIM] number, 180200), an embryonic malignant neoplasm of retinal origin that is associated with a low risk of osteosarcoma [30]. The expert panel considered the risk of retinoblastoma to be negligible for this patient, since he had reached the age of 8 years, but decided that it was important to inform the parents of the small chance that a sudden, painful swelling of the limbs could be caused by an osteosarcoma and that they should consult an oncologist at the first sign of such swelling. No further incidental findings were encountered.

In conclusion, our study shows that exome sequencing can be used as a diagnostic procedure for patients with severe intellectual disability of unknown cause. The diagnostic yield, which was 16% in our series, may increase with improvement in methods and the identification of additional genes associated with intellectual disability.

SUPPLEMENTARY TABLES AND FIGURES

Supplementary data is freely available online; http://www.nejm.org/action/showSupplements?doi=10.1056%2FNEJMoa1206524&viewType=Popup&viewClass=Supp

REFERENCES

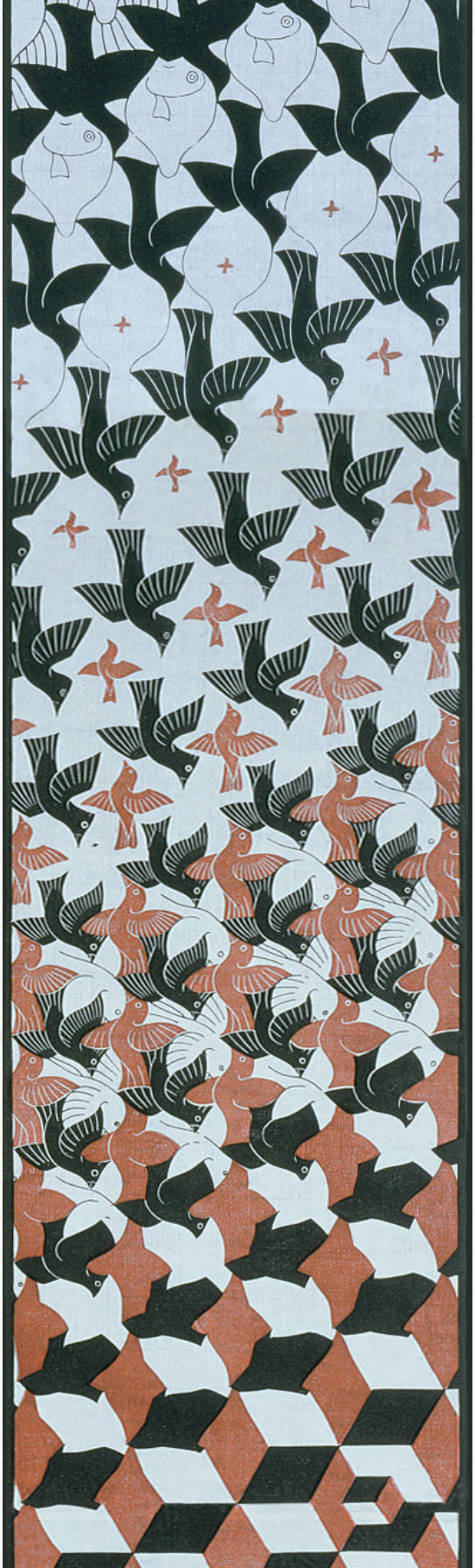
1. Leonard H, Wen X. The epidemiology of mental retardation: challenges and opportunities in the new millennium. Ment Retard Dev Disabil Res Rev 2002;8:117-34.
2. Ropers HH. Genetics of early onset cognitive impairment. Annu Rev Genomics Hum Genet 2010;11:161-87.
3. Topper S, Ober C, Das S. Exome sequencing and the genetics of intellectual disability. Clin Genet 2011;80:117-26.
4. Mefford HC, Batshaw ML, Hoffman EP. Genomics, intellectual disability, and autism. N Engl J Med 2012;366:733-43.
5. Battaglia A, Carey JC. Diagnostic evaluation of developmental delay/mental retardation: an overview. Am J Med Genet C Semin Med Genet 2003;117C:3-14.
6. Shea SE. Mental retardation in children ages 6 to 16. Semin Pediatr Neurol 2006;13:262-70.
7. Moeschler JB, Shevell M. Clinical genetic evaluation of the child with mental retardation or developmental delays. Pediatrics 2006;117:2304-16.
8. Romano C. The clinical evaluation of patients with mental retardation/intellectual

disability. In: Knight SJL, ed. Genetics of mental retardation. Vol. 18 of Monographs in human genetics. Basel, Switzerland: Karger, 2010:57-66.

9. Vissers LE, de Ligt J, Gilissen C, et al. A *de novo* paradigm for mental retardation. Nat Genet 2010;42:1109-12.
10. Hamdan FF, Gauthier J, Spiegelman D, et al. Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. N Engl J Med 2009;360:599-605.
11. Hamdan FF, Gauthier J, Araki Y, et al. Excess of *de novo* deleterious mutations in genes associated with glutamatergic systems in nonsyndromic intellectual disability. Am J Hum Genet 2011;88:306-16.
12. Sanders SJ, Murtha MT, Gupta AR, et al. *De novo* mutations revealed by whole exome sequencing are strongly associated with autism. Nature 2012;485:237-41.
13. O'Roak BJ, Vives L, Girirajan S, et al. Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. Nature 2012;485:246-50.
14. Iossifov I, Ronemus M, Levy D, et al. *De novo* gene disruptions in children on the autistic spectrum. Neuron 2012;74:285-99.
15. Veltman JA, Brunner HG. *De novo* mutations in human genetic disease. Nat Rev Genet 2012;13:565-75.
16. van Bokhoven H. Genetic and epigenetic networks in intellectual disabilities. Annu Rev Genet 2011;45:81-104.
17. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. N Engl J Med 2010;362:1181-91.
18. Najmabadi H, Hu H, Garshasbi M, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 2011;478:57-63.
19. Rauch A, Hoyer J, Guth S, et al. Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am J Med Genet A 2006;140:2063-74.
20. Bell J, Bodmer D, Sistermans E, Ramsden SC. Practice guidelines for the interpretation and reporting of unclassified variants (UVs) in clinical molecular genetics. Clinical Molecular Genetics Society, 2007 (http://www.cmgs.org/BPGs/pdfs%20current%20bpgs/UV%20GUIDELINES%20ratified.pdf).
21. Berg JS, Khoury MJ, Evans JP. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. Genet Med 2011;13:499-504.
22. Neale BM, Kou Y, Liu L, et al. Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. Nature 2012;485:242-5.
23. Willemsen MH, Vissers LE, Willemsen MA, et al. Mutations in DYNC1H1 cause severe intellectual disability with neuronal migration defects. J Med Genet 2012;49:179-83.
24. Debant A, Serra-Pagès C, Seipel K, et al. The multidomain protein Trio binds the LAR transmembrane tyrosine phosphatase, contains a protein kinase domain, and has separate rac-specific and rho-specific guanine nucleotide exchange factor domains. Proc Natl Acad Sci U S A 1996;93:5466-71.
25. Hochstenbach R, van Binsbergen E, Engelen J, et al. Array analysis and karyotyping:

workflow consequences based on a retrospective study of 36,325 patients with idiopathic developmental delay in the Netherlands. Eur J Med Genet 2009;52:161-9.

26. Sagoo GS, Butterworth AS, Sanderson S, Shaw-Smith C, Higgins JP, Burton H. Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. Genet Med 2009;11:139-46.
27. Miller DT, Adam MP, Aradhya S, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 2010;86:749-64.
28. Prasad C, Rupar T, Prasad AN. Pyruvate dehydrogenase deficiency and epilepsy. Brain Dev 2011;33:856-65.
29. Catarino CB, Liu JY, Liagkouras I, et al. Dravet syndrome as epileptic encephalopathy: evidence from long-term course and neuropathology. Brain 2011;134:2982-3010.
30. Matsunaga E. Retinoblastoma: mutational mosaicism or host resistance? Am J Med Genet 1981;8:375-87.



Chapter 4

Detection of Clinically Relevant Copy Number Variants with Whole-Exome Sequencing

Joep de Ligt[1]*, Philip M. Boone[2]*, Rolph Pfundt[1], Lisenka E.L.M. Vissers[1], Todd Richmond[3], Joel Geoghegan[3], Kathleen O'Moore[3], Nicole de Leeuw[1], Christine Shaw[2,3], Han G. Brunner[1], James R. Lupski[2,4,5], Joris A. Veltman[1] and Jayne Y. Hehir-Kwa[1]

1. Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Institute for Genetic and Metabolic Disease, Radboud University Medical Centre, Nijmegen 6500 HB, The Netherlands
2. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas
3. Roche NimbleGen, Madison, Wisconsin
4. Department of Pediatrics, Baylor College of Medicine, Houston, Texas
5. Texas Children's Hospital, Houston, Texas

*These authors contributed equally

Human Mutation 2013 Oct;34(10):1439-1448

ABSTRACT

Copy number variation (CNV) is a common source of genetic variation that has been implicated in many genomic disorders. This has resulted in the widespread application of genomic microarrays as a first-tier diagnostic tool for CNV detection. More recently, whole-exome sequencing (WES) has been proven successful for the detection of clinically relevant point mutations and small insertion-deletions exome wide. We evaluate the utility of short-read WES (SOLiD 5500xl) to detect clinically relevant CNVs in DNA from 10 patients with intellectual disability and compare these results to data from two independent high-resolution microarrays. Eleven of the 12 clinically relevant CNVs were detected via read-depth analysis of WES data; a heterozygous single-exon deletion remained undetected by all algorithms evaluated. Although the detection power of WES for small CNVs currently does not match that of high-resolution microarray platforms, we show that the majority (88%) of rare coding CNVs containing three or more exons are successfully identified by WES. These results show that the CNV detection resolution of WES is comparable to that of medium-resolution genomic microarrays commonly used as clinical assays. The combined detection of point mutations, indels, and CNVs makes WES a very attractive first-tier diagnostic test for genetically heterogeneous disorders.

INTRODUCTION

Whole-exome sequencing (WES) has revolutionized Mendelian disease gene identification by providing a powerful tool for exome-wide detection of single-nucleotide variants (SNVs) and small insertions and deletions (InDels) [Bainbridge et al., 2013; Bamshad et al., 2011; Gilissen et al., 2012; Hanchard et al., 2013; Ng et al., 2010; O'Roak et al., 2012]. In addition, WES is being introduced as a diagnostic procedure for genetically heterogeneous diseases in a number of laboratories [de Ligt et al., 2012; Hanchard et al., 2013; Rauch et al., 2012]. Structural variation such as copy number variants (CNVs), also contributes to these disorders [Cooper et al., 2011; Lupski, 2009; Stankiewicz and Lupski, 2010], and is currently not routinely assessed from WES data. The identification of CNVs, in addition to SNVs and InDels, would increase the versatility of WES as a genome-wide variant detection method in research and diagnostics. It would reduce the number of genomic assays required per patient to reach a diagnosis and create new possibilities to analyze the combined effects of SNVs and structural variation within an individual [Kurotaki et al., 2005].

Genomic microarray platforms based on either single-nucleotide polymorphisms (SNPs) or comparative genomic hybridization (CGH) have proven highly successful as a robust, high-throughput method for CNV detection [Boone et al., 2010; Pinkel et al., 1998; Schaaf et al., 2011; Vissers et al., 2003]. Advances in technology have resulted in an increase in the number of probes being included on a single array

from hundred thousands of probes (medium resolution) to millions of probes (high resolution), resulting in both increased detection power and accuracy. The implication of CNVs in a wide range of congenital disorders including intellectual disability (ID) and developmental delay, as well as later onset common diseases such as schizophrenia and autism, has resulted in the widespread application of genomic microarrays as a first-tier diagnostic tool [Lupski, 2012; Mefford and Eichler, 2009; Miller et al., 2010; Vissers et al., 2010]. The resolution to detect CNVs using genomic microarrays is strongly governed by the spacing and number of interrogating oligonucleotide probes, and the microarray design [Boone et al., 2010; Hehir-Kwa et al., 2007; Pinto et al., 2011]. However, intragenic CNVs remain beyond the detection limit of most clinical genomic microarray analysis, with the exception of custom microarray designs with enhanced exonic coverage for selected disease genes [Boone et al., 2010].

In contrast to most available genome-wide microarrays, WES specifically targets exonic regions and is mostly blinded to the remainder of the genome. The most widely applied massively parallel sequencing technologies sequence short reads (50-125 bp), either as fragments or as paired ends [Bamshad et al., 2011]. The most commonly applied methods for CNV detection in WES data are based on the analysis of the read depth, utilizing the number of fragments mapping within a genomic region as a measure of the amount of DNA present at the locus. This measure is used to determine a ratio between a test sample and reference samples [Haraksingh et al., 2011; Klambauer et al., 2012; Krummet al., 2012; Plagnol et al., 2012], and results in an estimation of copy number for a given genomic segment, similar to what is used for array based platforms. Read count data can, however, be distorted by the capture procedure used to isolate the coding portions of the genome and by inaccurate alignment of sequencing reads to the reference genome. For example, it is well documented that the percentage of Guanine and Cytosine nucleotides in the region significantly influences the binding affinity during capture and sequencing [Metzker, 2010]. In addition, the presence of low copy repeats can negatively influence alignment of sequence reads to the reference genome and thereby distort copy number estimations of a region [Teo et al., 2012].

To date, CNV detection in next generation sequencing data has been largely limited to sporadic cases and healthy control populations in a research setting [Mills et al., 2011]. Here, we evaluate the detection of clinically relevant, rare *de novo* CNVs of varying size and copy number state via WES. We compare the performance of WES for CNV detection with that of both commercially available as well as custom designed, high-resolution array CGH enhanced for coding regions using up to 4.2 million interrogating oligonucleotides.

MATERIALS AND METHODS

SAMPLE SELECTION

Ten samples were selected that had previously been diagnostically reported as containing at least one clinically relevant, rare *de novo* CNV associated with ID, detected by routine microarray based screening within the Department of Human Genetics, Radboud University Medical Centre, Nijmegen. These CNVs were chosen to represent a wide range of clinically relevant CNVs detected by microarray based analysis in our Genome Diagnostics division. The selected CNVs (1) contained at least one coding region, (2) were validated *de novo* using the same microarray platform on parental DNAs, (3) occurred across a variety of chromosomes, (4) ranged in copy number state from zero to three, and (5) ranged in genomic size from 15 kb to 24 Mb (**Table 1**). Eleven of these *de novo* CNVs were detected using an Affymetrix 250k NspI (Affymetrix, Santa Clara, CA) microarray and one, in patient 1, with the Affymetrix 2.7M microarray platform (**Table 1**).

WES AND CNV DETECTION

WES was performed as described by de Ligt et al. (2012); in brief, genomic DNA from these 10 samples was isolated from blood using the QIAamp DNA Mini Kit (Qiagen, Venlo, The Netherlands). Exomes were enriched using a SOLiD-Optimized Agilent SureSelect Human Exome Kit, V2 (Agilent Technologies, Santa Clara, CA), followed by SOLiD sequencing using a 5500xl System (Life Technologies, Carlsbad, CA) to a median read depth of 67 across targeted regions. Read correction and mapping were performed with Lifescope v1.3 (Life Technologies), using default settings. The WES data were analyzed with four different published CNV detection programs; (1) cn.MOPS v1.6.4 [Klambauer et al., 2012], (2) CONTRA v2.0.3 [Li et al., 2012], (3) CoNIFER v0.2.0 [Krumm et al., 2012], and (4) ExomeDepth v0.8.4 [Plagnol et al., 2012] (see Supplementary Methods), with unique hg19-based RefSeq gene exon definitions as target regions in the analysis.

ADDITIONAL GENOMIC MICROARRAY STUDIES

All samples were also analyzed on two independent, microarray platforms: (1) a high-resolution SNP microarray (Affymetrix CytoScanHD with 2.6 million probes; “CytoScanHD”) (Affymetrix) and (2) a high-density CGH microarray enhanced for exonic regions (NimbleGen 4.2 million probe custom design; “ExonArray”) (Roche NimbleGen, Madison, WI). Detailed experimental methods and computational approaches/software parameters are described in the Supplementary Methods.

The aim of the ExonArray design was to cover each exon (Supplementary Methods), and flanking sequence, with at least eight oligonucleotide probes. After testing and optimization (see Supplementary Methods), the ideal coverage of eight or more

probes was achieved for over 135,000 (~86%) exons; 249 (0.16%) of the exons could not be targeted at all. To test the sensitivity of the ExonArray, seven DNAs with 10 previously described CNVs (nine deletions and one duplication) with a median size of 8.5 kb (size range 1.6 kb-1.7 Mb) [Boone et al., 2013; Zhang et al., 2009] were analyzed (Supplementary Figure S1). NimbleGen performed the microarray experiments in a blinded fashion using mixed control DNAs. All the 10 CNVs were identified successfully indicating 100% sensitivity for these events, which were as small as 1.6 kb, five being smaller than 10 kb, and of which four encompassed only a single exon (Supplementary Figure S1).

CNV ANNOTATION

Prior to annotation and interpretation, CNV calls resulting from both the WES approach and the ExonArray were subject to additional merging (Supplementary Methods). To facilitate interpretation, we annotated all CNVs for their gene content, (UCSC hg19 track GeneSymbols), the total number of genes, and the number of unique coding exons within the region. Since mapping artifacts can lead to false positive (FP) signals in sequencing data, the CNVs were annotated for features related to the uniqueness of the genomic region, the repeat content (simple and complex), and the percentage of SelfChain alignment in the region, based on the UCSC repeat tracks.

A reference set was generated to represent common CNV regions detected by the different platforms (both high-resolution microarrays and WES) and algorithms used in this study to determine which genomic regions were copy number variable (common CNVs). The reference set contained all events observed in more than one individual, by any specific platform in this study, as well as CNVs identified in our in-house set of control samples. This in-house dataset contains CNVs identified in 1,200 healthy individuals analyzed with the Affymetrix 6.0 SNP microarray platform [Franke et al., 2010] and 650 individuals analyzed with the Affymetrix CytoScanHD. The combined dataset included in total 23,125 gains and 56,066 losses.

OVERALL CNV DETECTION POWER OF WES

The false negative (FN) detection rate of WES was calculated by measuring the number of CNV events detected using the high resolution microarray platforms that were missed by WES. To prevent overestimation due to platform design (exon targeted vs. whole genome), we accounted for both the exome enrichment targets and the detection power of WES. We selected CNVs that were identified by at least two independent microarray platforms (minimum overlap of 30% of the CNV region, to allow for breakpoint inaccuracies due to the large differences in probe densities) and the CNV had to encompass at least three exons. For each

Table 1. *Overview of the Detection of 12 Clinically Relevant De Novo CNVs*

		Discovery microarray					WES read-depth algorithms			
Patient	**Chromo-some**	**Estimated start position (kb)**	**Estimated end position (kb)**	**CNV size (kb)**	**Copy number state**	**Nr. Genes**	**CONTRA**	**cn.MOPS**	**Exome-Depth**	**CoNIFER**
1	chr10	89,642.6	89,657.5	14.9	1	1[a]	-	-	-	-
2	chr19	33,371.1	33,394.2	23.0	0	1	-	-	V	V
3	chr8	77,745.6	77,795.2	49.6	1	1	-	-	V	V
4	chr17	1,203.6	1,516.5	312.9	3	8	-	-	V	V
5	chr16	29,673.2	29,988.3	315.1	1	16	-	-	V	V
6	chr15	43,759.8	44,862.9	1,103.2	1	24	-	-	-	V
7	chr2	233,166.3	233,886.7	720.5	3	16	-	-	V	V
8	chrX	6,495.3	7,951.7	1,456.4	0	5	-	-	V	V
9	chr2	239,952.7	241,373.1	1,420.5	3	14	-	-	V	V
	chr2	241,442.7	243,001.9	1,559.2	1	31	-	-	V	V
	chr15	60,489.7	62,906.5	24,603.6	3	210	-	-	V	V
10	chr20	77,771.0	102,374.6	2,416.8	3	91	-	V	V	V

CNVs as detected by the discovery microarray (hg19), genomic location, size, predicted copy number state and the number of genes in the region.

a.) A single exon deletion. Detection by the different WES approaches; –, CNV is not detected with a minimum overlap of 30%, V, detected with a minimum overlap of 30%.

CNV, the largest region, detected by the CytoScanHD or the ExonArray, was used for further analysis. After applying these selection criteria to the total set of 6,074 CNV identified by the different microarray experiments, the resulting consensus dataset contained 97 CNVs. Of these 97 consensus CNVs, 25 did not occur in the common CNV dataset and were considered rare CNVs. Consensus CNVs were only considered as positively detected by WES if a CNV was called in the same region and overlapped the consensus CNV region for at least 30%.

BREAKPOINT ANALYSIS

To study the differences in detected CNV breakpoints across detection platforms, an overlap analysis was performed on the 11 clinically relevant CNVs. CNVs overlapping the discovery region were merged into a maximum confirmation CNV, and breakpoint differences were calculated based on the genomic coordinates of the two CNVs. The difference in genomic location was measured for each breakpoint by subtracting the genomic location as defined by the high-resolution array consensus from the location identified by the confirmation platform.

DATA AVAILABILITY

CNVs identified in this study by the different platforms have been submitted to dbVar under nstd84; sample identifiers correspond to those used in this paper. Detailed information on clinical presentation and the pathogenic event is available through ECARUCA for all patients under the following accession numbers (patient 1-10): 5042, 5045, 4785, 5044, 4545, 4487, 4581, 5043, 4452, and 4685, respectively. Raw data of the discovery microarray experiments are available in the Gene Expression Omnibus (GSE46060); sample identifiers correspond to those used in this paper.

RESULTS

Our study aimed to investigate the diagnostic potential of CNV identification from short-read WES (SOLiD 5500xl) data. For this, we selected a set of 12 clinically relevant and validated, rare *de novo* CNVs detected using either an Affymetrix 250k NspI or 2.7M microarray, in 10 individuals with ID. This set of CNVs varied in genomic size and copy number state and incorporated both autosomal and X-linked CNVs (**Table 1**). WES was performed on all 10 samples and CNVs were called using four published CNV detection algorithms. In addition, high-resolution microarray experiments were performed using two independent platforms to experimentally assess the genome-wide true positive (TP), FP, and FN CNV detection rates of WES.

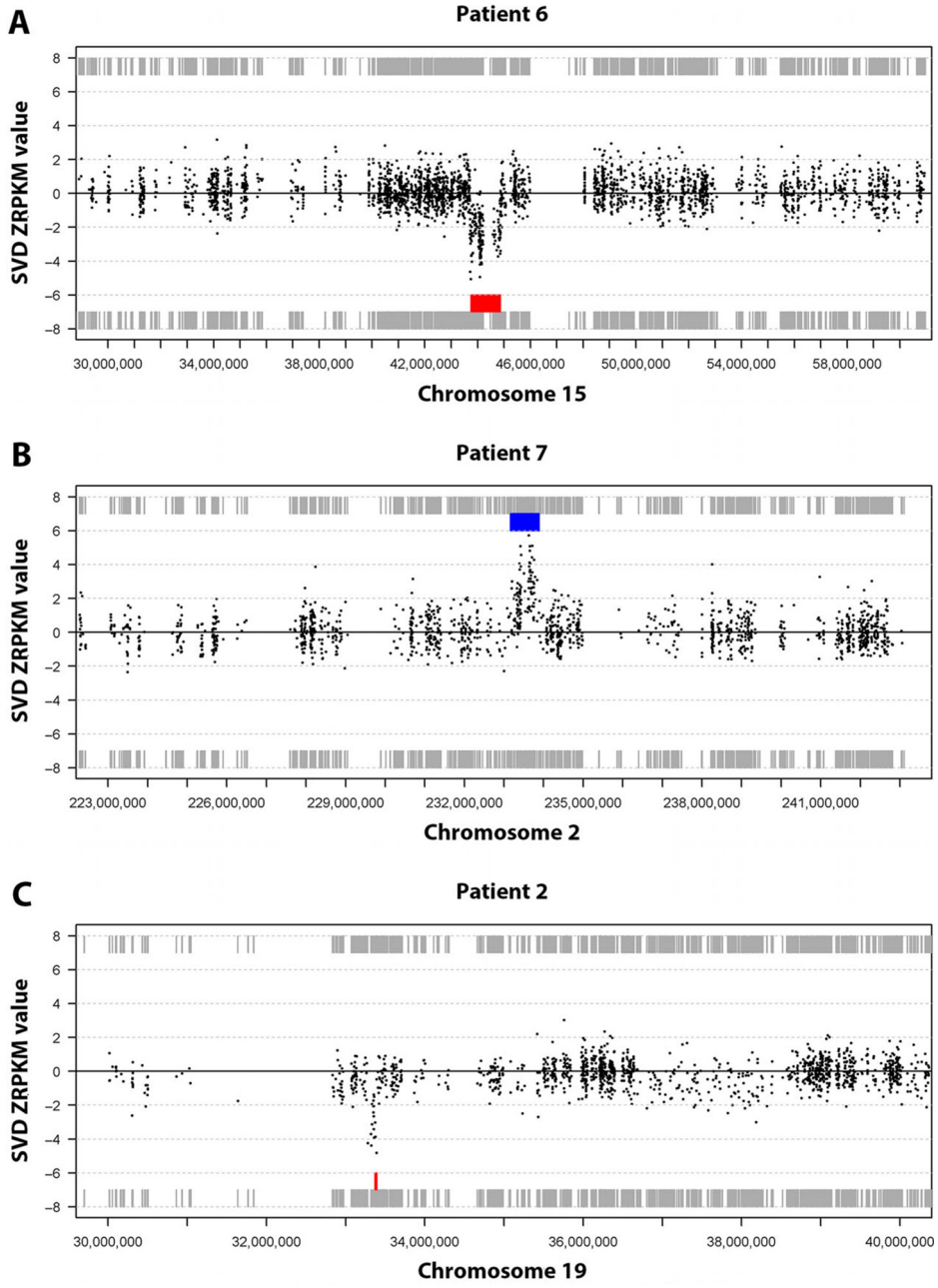


Figure 1. *Detection of clinically relevant CNVs by WES. Black circles represent test over reference ratio values generated from WES data using CoNIFER; singular value decomposition (SVD) and Z-score adjusted read count per million (ZRPKM). Gray boxes indicate RefSeq gene exons; boxes above the ratio values represent CNV gains and below CNV deletions. A: A 1.1 Mb deletion including 24 genes. B: A 720 kb duplication including 17 genes. C: A 23 kb deletion containing two genes.*

Table 2. *Performance of WES CNV Detection Algorithms*

Algorithm	Array confirmation of all WES CNVs	Array confirmation of rare WES CNVs
cn.MOPS (n=329)	163 (49.5%)	33 (44.7%)
CoNIFER (n=65)	38 (58.5%)	28 (53.9%)
Contra (n=1,464)	248 (16.9%)	75 (11.0%)
ExomeDepth (n=1,482)	234 (15.8%)	48 (6.6%)

Algorithm	Failed to confirm by WES of all array consensus CNVs	Failed to confirm by WES of rare array consensus CNVs
cn.MOPS	109 (89%)	32 (97%)
CoNIFER	91 (74%)	8 (24%)
Contra	123 (100%)	33 (100%)
ExomeDepth	89 (72%)	23 (45%)

CoNIFER has the highest TP rate and the lowest FN rate both for rare CNVs across the 10 patients. The TP rate measures the number of CNVs identified by each WES-based algorithm that were confirmed by a CNV called on a high-resolution microarray platform. The FN rate is calculated by creating a consensus of CNVs identified with the microarray platforms (Materials and Methods) and determining the number of CNVs not reliably detected by each WES-based algorithms. The best performing algorithm is highlighted in bold.

DETECTION OF THE CLINICALLY RELEVANT CNVS

The four different WES CNV identification algorithms varied in their ability to correctly identify the 12 clinically relevant CNVs (**Table 1**). ExomeDepth and CoNIFER performed best, correctly identifying 10 and 11 of the 12 clinically relevant CNVs, respectively (**Table 1**, **Figure 1**, for examples of WES-based CNV detection using CoNIFER). Of note, all WES algorithms failed to detect a clinically relevant single exon deletion (15 kb in size) in patient 1, which was originally detected using the Affymetrix 2.7M microarray. While CONTRA and cn.MOPS often called a CNV in the relevant CNV region, the identified CNV was small and overlapped less than 30% (cut off threshold used for successful detection) with the interval identified by the discovery microarray. The copy number state reported by the WES-based CNV algorithms matched the copy number estimated by the microarrays for all CNVs.

GENOME-WIDE CNV DETECTION USING WES

The four different CNV WES detection algorithms varied widely in the total number of CNVs detected across the 10 samples; CONTRA identified 1,464 CNVs, ExomeDepth 1,482 CNVs, cn.MOPS 329 CNVs, and CoNIFER 65 CNVs in total (Supplementary Table S1). All but one (99.9%) of the CNV events identified by CONTRA contained three or fewer coding exons. Similarly, many CNVs identified by cn.MOPS (56%) and ExomeDepth (58%) also contained three or fewer coding exons; in contrast, CoNIFER focuses more on detecting larger and rare CNVs, and detected only six (9%) such small CNVs (Supplementary Figure S2).

To evaluate the reliability of CNV identification using WES, we compared the results to CNVs detected by the different microarray platforms used in this study (Affymetrix 250k NspI/2.7M, Affymetrix CytoScanHD and the NimbleGen 4.2M ExonArray). In total, 38 of the 65 (59%) CNVs identified using CoNIFER were supported by one (n=10) or more (n=28) of the microarray platforms (**Table 2**). The confirmed events were larger (median 63.7 kb) and contained more exons (median 21.5) compared to the unsupported CNVs (median size 16.3 kb, median number of exons is 7). Similarly, 50% of the CNVs identified by cn.MOPS were supported by a microarray CNV, whereas a much smaller proportion of CNVs identified by CONTRA (17%) and ExomeDepth (16%) was supported by one or more microarray platforms (**Table 2**).

While genome-wide accuracy measures are an important indication of algorithm performance, in a clinical setting, it is important to consider the number of missed rare, genic events. To evaluate the FN rate of WES CNV detection, we investigated the detection of a CNV consensus set containing 25 rare coding CNVs detected by the two highest resolution microarray platforms used (ExonArray and CytoscanHD, see Materials and Methods). The overlap analysis showed the best detection rate for CoNIFER (88%), missing 3 of the 25 rare, genic events (**Table 2**). In general, the

Table 3. *Breakpoint Accuracy of WES CNV Detection*

		High resolution microarray consensus				WES detection [a]			
Patient	**Chromo-some**	**Start position (deviation) (kb)**	**Stop position (deviation) (kb)**	**Size (kb)**	**Nr. Genes min-max**	**Start difference (kb)**	**Stop difference (kb)**	**Size difference (kb)**	**Nr. Genes**
1	chr10	x	x	x	x	x	x	x	-
2	chr19	33,352 (17)	33,396 (1)	44	2	-88.6	13.0	101.5	3
3	chr8	77,720 (0.4)	77,808 (0.3)	88	1	-100.6	1770.2	1870.8	4
4	chr17	1,142 (2.2)	1,494 (2.7)	351	9	-147.2	24.8	172.0	10
5	chr16	29,640 (11.6)	30,189 (11.3)	548	27-28	-92.9	10.9	103.7	30
6	chr15	43,714 (1.6)	44,863 (0.1)	1,150	24	-5.9	2.6	8.5	24
7	chr2	233,149 (1.1)	233,898 (2.0)	749	16-17	-34.7	1.7	36.4	17
8	chrX	6,453 (3.4)	8,001 (1.1)	1,548	5-6	-626.0	500.4	1126.4	12
9	chr2	239,945 (1.5)	241,423 (0.5)	1,478	18	71.6	-15.3	-86.9	15
	chr2	241,428 (5.0)	242,918 (134.5)	1,490	29-32	11.1	139.0	127.9	32
	chr15	77,765 (1.1)	102,415 (13.9)	24,650	200-211	5.7	80.3	74.6	213
10	chr20	60,463 (0)	62,940 (24.6)	2,478	91-92	1602.4	-40.8	-1643.2	43
Median difference (stdev)						88.5 (313)	24.7 (326)	103.7 (578)	3 (8)

The difference in breakpoints for the 11 clinically relevant CNVs detected on all platforms, compared with the average breakpoint positions detected by the two highest resolution platforms (CytoScanHD and ExonArray).
a) WES CNV detection by CoNIFER. Nr. genes, the number of genes based on RefSeq gene definitions (UCSC hg19); differences in breakpoint positions = WES – high-resolution microarray consensus; + indicates a higher genomic position; – indicates a lower genomic position. For size differences, – indicates undercalling of the CNV size by WES. stdev, standard deviation.

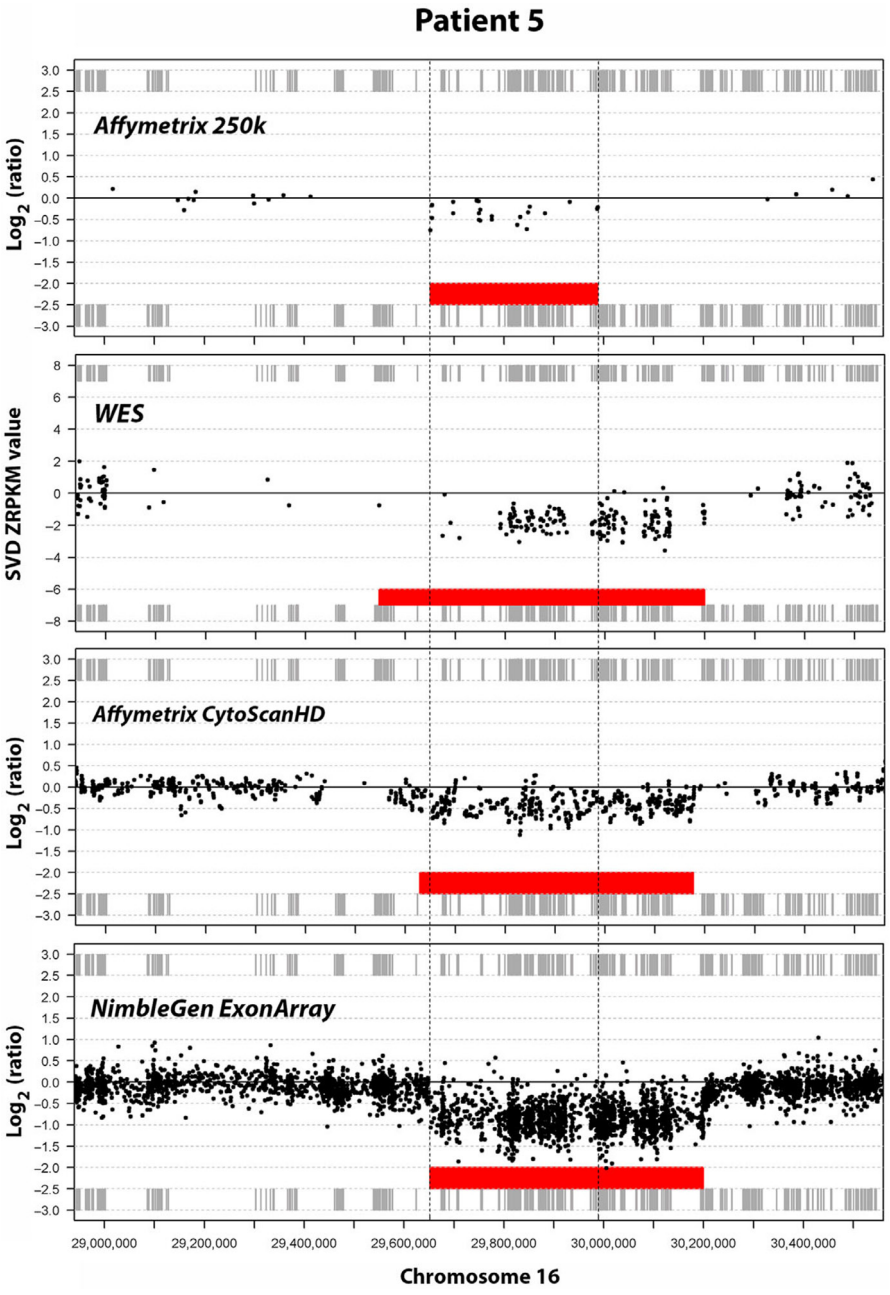
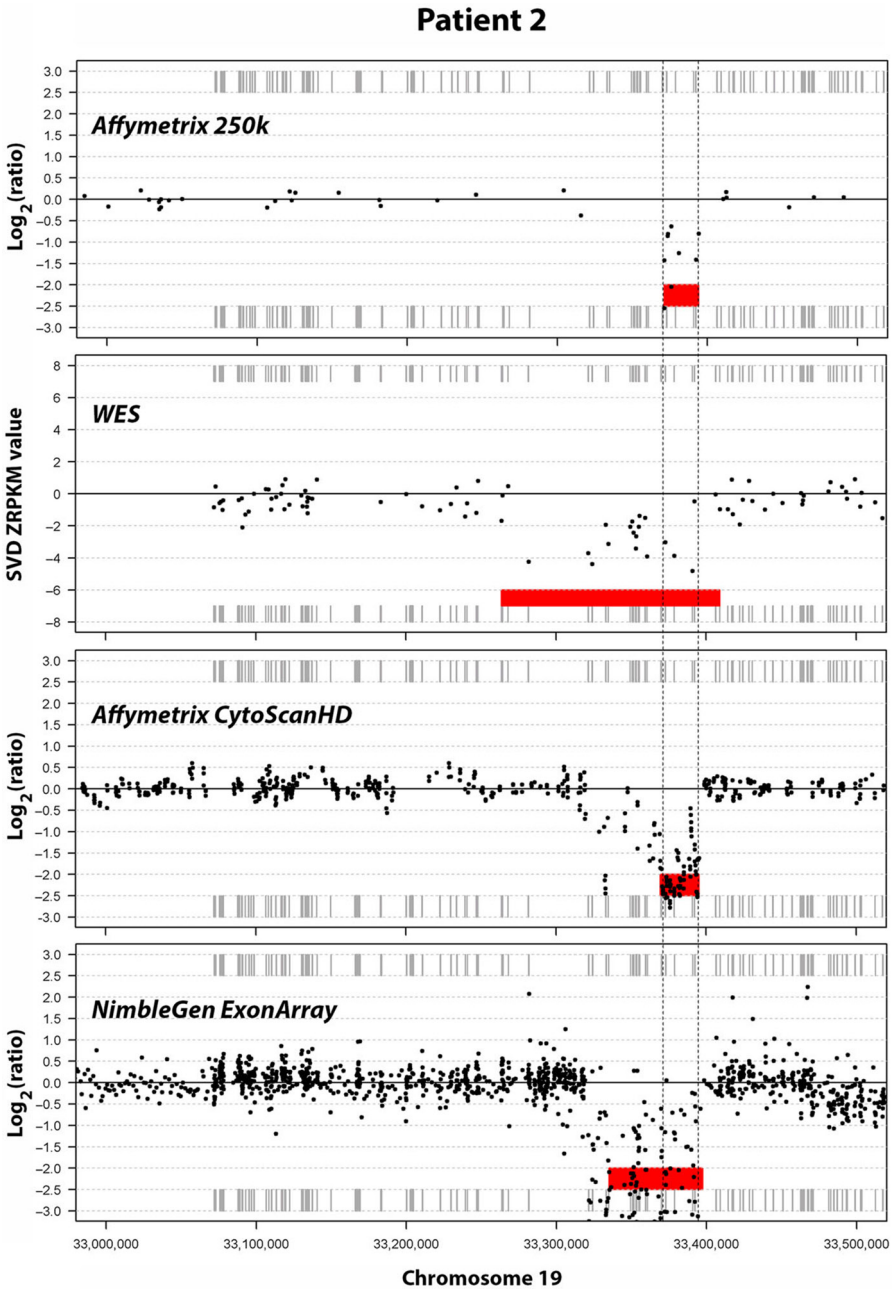


Figure 2. *Detection of clinically relevant CNVs in two patients by the different platforms. Black circles are the log2 test over reference ratio values obtained through the different microarray experiments and singular value decomposition (SVD) and Z-score adjusted read count per million (ZRPKM) values for WES. (legend continues on the next page)*



Boxes below ratio values represent the CNV deletion as detected by the different platforms; gray boxes indicate RefSeq gene exons, and the vertical dotted lines indicate the minimally deleted region detected by the discovery microarray.

CNV events not detected by the WES approach using CoNIFER contained fewer exons (3, 4, and 17 exons) than those CNVs that were detected (median 27 exons). Overall, CoNIFER proved to be the most reliable CNV detection program for diagnostic applications based on a number of features: 1) most of the clinically relevant CNVs were detected (11 out of 12); 2) the highest percentage of CNVs were supported by an independent method (59%); and 3) the lowest number of rare consensus CNVs were missed (12%). All of the other algorithms had high FP rates and missed large numbers of rare CNVs, as well as the clinically relevant CNVs (**Tables 1 & 2**), making them unsuitable for diagnostic applications. Further analyses of WES CNV identification were based on the calls by CoNIFER.

DETERMINING THE ACCURACY OF CNV IDENTIFICATION

The experimental design of WES leads to a nonuniform distribution of data points, focused only on the coding regions, whereas most genomic microarrays have a probe distribution containing a backbone covering the entire genome. To assess the effect of the unequal probe distribution of WES on the accuracy of CNV identification, we compared the breakpoints of the 11 clinically relevant CNVs detected across all four experimental platforms (discovery microarray, WES, CytoScanHD, and the ExonArray).

We generated consensus breakpoints based on the results from the highest resolution microarray platforms (CytoScanHD and ExonArray). The breakpoints of the WES CNV detection mapped within 200 kb of the consensus breakpoints for 18 of the 22 breakpoints (**Table 3**, **Figure 2**, for example plots of CNV detection on different platforms). Three of the four breakpoints, which deviated more than 200 kb, occurred in regions with a much lower exon density as compared with the well mapped breakpoints (mean of 1.83 vs. 59.5 exons within 500 kb of the breakpoint region).

PREDICTING THE DIAGNOSTIC YIELD OF CNV IDENTIFICATION USING WES

After determining the power to detect CNVs in WES data, we estimated the impact of using WES CNV detection on a larger set of samples within a clinical setting. For this, we compiled a list of 470 clinically relevant *de novo* CNVs detected by diagnostic microarray analysis in our center using a combination of Affymetrix 250k NspI, 2.7M and CytoScanHD microarray platforms. For each *de novo* CNV, the number of exons present in the sequencing capture set was calculated to determine if the CNV could be detected via WES. In total, 97% of the CNVs contained three or more exons, the minimum number required for WES-based CNV calling by CoNIFER. The majority of CNVs in this larger set of clinically relevant CNVs were larger than 200 kb in size (96%), whereas only half of CNVs from the rare consensus set (13/25) were in this

size range. WES achieved a detection rate of 75% for CNVs smaller than 200 kb, and 100% for CNVs larger than 200 kb. When we apply these detection rates to the clinically relevant CNVs, it is predicted that 96% (453 CNVs) of these CNVs would have been successfully identified by WES. Based on the limited number of CNVs included in this study, this theoretical detection rate is in line with the observed experimental detection rate of 92% (i.e., 11 of the 12 clinically relevant CNVs being successfully detected).

DISCUSSION

CNV is a common source of genetic variation that has been implicated in many genomic disorders [Cooper et al., 2011; Lupski, 2009; Stankiewicz and Lupski, 2010]. This has resulted in the widespread application of genomic microarrays as a first-tier diagnostic tool for CNV detection [Mefford and Eichler, 2009; Miller et al., 2010; Stankiewicz and Beaudet, 2007; Vissers et al., 2010]. The introduction of massive parallel sequencing approaches has provided a valuable tool for mutation identification in rare and genetically heterogeneous disorders [Bamshad et al., 2011; de Ligt et al., 2012; Gilissen et al., 2012; Gonzaga-Jauregui et al., 2012; Hanchard et al., 2013; Ng et al., 2010; O'Roak et al., 2012; Rauch et al., 2012]. For example, in a genetically heterogeneous disorder such as ID, a causal or candidate (*de novo*) mutation was identified in up to 38% of cases [de Ligt et al., 2012; Rauch et al., 2012], and it has been reported that an additional 10%-20% of ID cases can be explained by clinically relevant *de novo* CNVs [Mefford et al., 2012]. Therefore, the addition of CNV detection from WES data could achieve a diagnostic yield up to 58%, with a single test, for ID. This would represent the highest diagnostic yield of any current clinical genetic screening method for this disorder. A single genomic assay, which detects all forms of genomic variation, could decrease the time to obtain a molecular diagnosis, and reduce the diagnostic odyssey faced by patients and families.

Here, we evaluated the utility of WES to detect known clinically relevant CNVs in 10 patients. We tested four different CNV detection algorithms for WES data and compared their results to CNVs detected by three different genomic microarray platforms. These results provide insights into the possibilities and limitations of CNV detection using different experimental platforms currently available, as well as the performance of CNV identification algorithms with both WES data and high-resolution genomic microarrays.

Of the four algorithms tested in this study, CoNIFER was found to perform best with the highest TP rate and the lowest FN rate for the detection of rare coding CNVs. It is likely that CoNIFER performs especially well for rare CNVs due to the rigorous

correction for systematic fluctuation, as well as enrichment, sequencing, and mapping biases, by singular value decomposition and the use of a Z-score approach, which corrects for positional fluctuation across samples. The FN rate of CoNIFER was greater for common CNVs, which is likely due to the Z-score approach applied for copy number estimation. The Z-score corrects for the fluctuation of a data point in the reference set; as a result, CNVs occurring in a region where reference samples are variable will have a lower Z-score compared with the same CNVs in a copy number stable region. Additionally, we limited our analysis to read algorithms suitable for short (50 bp) single-end reads as sequenced by SOLiD chemistry because read-depth algorithms are applicable to most WES approaches [Bamshad et al., 2011]. When longer reads or read pairs are available, more sophisticated methods can be used to increase the detection power for CNVs by combining different lines of evidence such as split read and clustering of discordant pairs [Teo et al., 2012] with a wider range of available programs [Duan et al., 2013].

Identifying CNVs in WES data is subject to a number of limitations due to the uneven spacing of exons, and thus data points, across the genome [Teo et al., 2012]. This affected the identification of the CNV segments, which in four cases were over segmented and reported as several smaller CNVs, requiring merging during post-processing. Likewise, the unequal spacing of the genomic data points also influenced the identification of the CNV breakpoints.

In general, the maximum possible size of the CNVs was reported; and in the absence of data points, segments were continued until a normal copy number signal was detected. Alternatively, CNV breakpoints can be identified based on the last occurrence of an aberrant copy number signal, the minimum CNV size. The difference between the maximum and minimum predicted CNV size as called by WES varied between 2.8 and 542.8 kb across the 11 *de novo* CNVs. Reporting both the maximum and minimum possible CNVsize provides useful insights into the uncertainty of breakpoint predictions.

Most clinically relevant CNVs currently detected by routine screening are large (Supplementary Figure S2) and often contain multiple genes. Likewise, the CNVs identified in this study using WES were biased to larger CNVs containing multiple exons. However, our results using high-resolution microarrays indicate a large number of smaller single exon CNVs may exist within these samples (Supplementary Table S2). Likewise, data from personal genomes [Wheeler et al., 2008], high resolution CGH arrays [Conrad et al., 2010], and WES [Mills et al., 2011] indicate that the genomes of healthy individuals harbor 600-900 [Korbel et al., 2008; Levy et al., 2007] CNVs with a median size of 0.7 kb. Validation experiments of the 4.2M NimbleGen microarray (ExonArray) showed that this platform has the potential to reliably detect known single exon deletions, and screening for exon level CNVs in a

clinical setting has revealed multiple small, causal events [Boone et al., 2010;Whibley et al., 2010]. These small CNV events have been largely invisible to commercial genome-wide microarrays and remain challenging to detect through WES. While the detection specificity and sensitivity of the platforms used in this study is unclear for these small CNVs, it is apparent that these events occur frequently and could contribute to the patient's phenotype. Thorough validation studies of these very small CNVs are required to establish their frequency and possible contribution to disease.

While the current detection power of WES, especially for single exon CNVs, does not match that of high-resolution microarray platforms, we show that WES data are suited for the detection of large, rare, genic events that represent the majority of currently reported clinically relevant CNVs. A likely reason why the single exon 15 kb deletion included in this study was difficult to detect is that each exon represents a single data point. Detecting the difference between signal and experimental noise based on one data point requires very little fluctuation or noise. Possible solutions for larger exons include subdivision of exons into smaller regions to create multiple data points, or in the case of deletions, to include homozygosity data from SNVs into the detection algorithm. Ongoing developments in CNV identification algorithms will likely result in further performance improvements [Amarasinghe et al., 2013; Fromer et al., 2012].

The reliable detection of rare, genic CNVs is a valuable adjuvant tool within the clinical setting when WES data are available. Possibilities to enhance the detection power for CNVs of WES approaches include larger capture kits, the addition of a genomic backbone to improve genome-wide resolution, and/or the addition of intronic capture sequences to improve the accuracy in determining which exons are affected by a CNV. Improvements in data analysis could be made by applying more sophisticated normalization methods to account for biases introduced during the capture and sequencing procedures. In addition, current WES CNV detection algorithms used in this study are limited in breakpoint accuracy by the read-depth approach and could be further improved by incorporating information from genotypes, split-reads, and read-pair information to increase the detection power of WES for CNVs [Mills et al., 2011]. While these improvements are of great benefit to further increase WES-based CNV detection, the results presented in this study show that CNV detection resolution of exome sequencing is already comparable to that of medium-resolution genomic microarrays currently used as clinical assays.

ACKNOWLEDGMENTS

The authors wish to thank Dorien Lugtenberg for her useful discussions and input during this study. We also thank the Genome Diagnostics group at UMCN for performing the Affymetrix CytoScanHD experiments and NimbleGen for their support in the validation experiments of the ExonArray.

SUPPLEMENTARY TABLES AND FIGURES

Supplementary data is freely available online; http://onlinelibrary.wiley.com/doi/10.1002/humu.22387/suppinfo

REFERENCES

Amarasinghe KC, Li J, Halgamuge SK. 2013. CoNVEX: copy number variation estimation in exome sequencing data using HMM. BMC Bioinformatics 14 Suppl 2:S2.

Bainbridge MN, Hu H, Muzny DM, Musante L, Lupski JR, Graham BH, Chen W, Gripp KW, Jenny K, Wienker TF, Yang Y, Sutton VR, et al. 2013. *De novo* truncating mutations in ASXL3 are associatedwith a novel clinical phenotypewith similarities to Bohring-Opitz syndrome. Gen Med 5:11.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12:745-755.

Boone PM, Bacino CA, Shaw CA, Eng PA, Hixson PM, Pursley AN, Kang SH, Yang Y, Wiszniewska J, Nowakowska BA, del Gaudio D, Xia Z, et al. 2010. Detection of clinically relevant exonic copy-number changes by arrayCGH. Hum Mutat 31:1326-1342.

Boone PM, Soens ZT, Campbell IM, Stankiewicz P, Cheung SW, Patel A, Beaudet AL, Plon SE, Shaw CA, McGuire AL, Lupski JR. 2013. Incidental copy-number variants identified by routine genome testing in a clinical population. Genet Med 15: 45-54.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704-712.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, et al. 2011. A copy number variation morbidity map of developmental delay. Nat Genet 43:838-846.

de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, et al. 2012. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367:1921-1929.

Duan J, Zhang JG, Deng HW, Wang YP. 2013. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. PLoS One 8:e59128.

Franke B, Vasquez AA, Veltman JA, Brunner HG, Rijpkema M, Fernández G. 2010. Genetic variation in *CACNA1C*, a gene associatedwith bipolar disorder, influences brainstem rather than gray matter volume in healthy individuals. Biol Psychiatry 68:586-588.

Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF, et al. 2012. Discovery and statistical genotyping of copy-number variation fromwhole-exome sequencing depth. Am J Hum Genet 91:597-607.

Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20:490-497.

Gonzaga-Jauregui C, Lupski JR, Gibbs RA. 2012. Human genome sequencing in health and disease. Annu Rev Med 63:35-61.

Hanchard NA, Murdock DR, Magoulas PI, Bainbridge M, Muzny D, Wu Y, Wang M, McGuire AL, Lupski JR, Gibbs RA, Brown CW. 2013. Exploring the utility of whole-exome sequencing as a diagnostic tool in a child with atypical episodic muscle weakness. Clin Genet 83:457-461.

Haraksingh RR, Abyzov A, Gerstein M, Urban AE, Snyder M. 2011. Genome-wide mapping of copy number variation in humans: comparative analysis of high-resolution array platforms. PLoS One 6:e27859.

Hehir-Kwa JY, Egmont-Petersen M, Janssen IM, Smeets D, van Kessel AG, Veltman JA. 2007. Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. DNA Res 14:1-11.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 40:e69.

Korbel JO, Kim PM, Chen X, Urban AE, Weissman S, Snyder M, Gerstein MB. 2008. The current excitement about copy-number variation: how it relates to gene duplications and protein families. Curr Opin Struct Biol 18:366-374.

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson DA, Eichler EE. 2012. Copy number variation detection and genotyping from exome sequence data. Genome Res 22:1525-1532.

Kurotaki N, Shen JJ, Touyama M, Kondoh T, Visser R, Ozaki T, Nishimoto J, Shiihara T, Uetake K, Makita Y, Harada N, Raskin S, et al. 2005. Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating coagulation factor twelve (FXII) deficiency. Genet Med 7:479-483.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, et al. 2007. The diploid genome sequence of an individual human. PLoS Biol 5:e254.

Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL. 2012. CONTRA: copy number analysis for targeted resequencing. Bioinformatics 28:1307-1313.

Lupski JR. 2009. Genomic disorders ten years on. Genome Med 1:42.

Lupski JR. 2012. Brain copy number variants and neuropsychiatric traits. Biol Psychiatry 72:617-619.

Mefford HC, Batshaw ML, Hoffman EP. 2012. Genomics, intellectual disability, and autism. N Engl J Med 366:733-743.

Mefford HC, Eichler EE. 2009. Duplication hotspots, rare genomic disorders, and common disease. Curr Opin Genet Dev 19:196-204.

Metzker ML. 2010. Sequencing technologies the next generation. Nat Rev Genet 11:31-46.

Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, et al. 2010. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 86:749-764.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470:59-65.

Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S. 2005. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 65:6071-6079.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30-35.

O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, Munson J, Hiatt JB, et al. 2012. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science 338:1619-1622.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207-211.

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, et al. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 29:512-520.

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, Kumararatne D, Doffinger R, et al. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28:2747-2754.

Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, Dufke A, Cremer K, et al. 2012. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 6736:1674-1682.

Schaaf CP, Wiszniewska J, Beaudet AL. 2011. Copy number and SNP arrays in clinical diagnostics. Annu Rev Genomics Hum Genet 12:25-51.

Stankiewicz P, Beaudet AL. 2007. Use of arrayCGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. Curr Opin Genet Dev 17:182-192.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. Annu Rev Med 61:437-455.

Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics 28:2711-2718.

Vissers LELM, de Vries BBA, Osoegawa K, Janssen IM, Feuth T, Choy CO, Straatman H, van der Vliet W, Huys EHLPG, van Rijk A, Smeets D, van Ravenswaaij-Arts CMA, et al. 2003. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities.AmJHum Genet 73:1261-1270.

Vissers LELM, de Vries BBA, Veltman JA. 2010. Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis. J Med Genet 47:289-297.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-876.

Whibley AC, Plagnol V, Tarpey PS, Abidi F, Fullston T, Choma MK, Boucher CA, Shepherd L, Willatt L, Parkin G, Smith R, Futreal PA, et al. 2010. Fine-scale survey of X chromosome copy number variants and indels underlying intellectual disability. Am J Hum Genet 87:173-188.

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. 2009. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 41:849-853.

Chapter 5

Non-invasive prenatal diagnosis using massively parallel sequencing

Publication title: Non-invasive prenatal diagnosis of fetal aneuploidies using massively parallel sequencing-by-ligation and evidence that cell-free fetal DNA in the maternal plasma originates from cytotrophoblastic cells

Brigitte HW Faas[1], Joep de Ligt, Irene Janssen, Alex J. Eggink, Lia D.E. Wijnberger, John M.G. van Vugt, Lisenka E.L.M. Vissers, Ad Geurts van Kessel

1. Radboud University Nijmegen Medical Centre, Department of Human Genetics, Nijmegen, The Netherlands

Expert Opinion on Biological Therapy. 2012 Jun;12 Suppl 1:S19-26

ABSTRACT

Blood plasma of pregnant women contains circulating cell-free fetal DNA (ccffDNA), originating from the placenta. The use of this DNA for noninvasive detection of fetal aneuploidies using massively parallel sequencing (MPS)-by-synthesis has been proven previously. Sequence performance may, however, depend on the MPS platform and therefore we have explored the possibility for multiplex MPS-by-ligation, using the Applied Biosystems SOLiD 4 system. DNA isolated from plasma samples from 52 pregnant women, carrying normal or aneuploid fetuses, was sequenced in multiplex runs of 4, 8 or 16 samples simultaneously. The sequence reads were mapped to the human reference genome and quantified according to their genomic location. In case of a fetal aneuploidy, the number of reads of the aberrant chromosome is expected to be higher or lower than in normal reference samples. To statistically determine this, Z-scores per chromosome were calculated as described previously, with thresholds for aneuploidies set at > +3.0 and < -3.0 for chromosomal over- or under representation, respectively. All samples from fetal aneuploidies yielded Z-scores outside the thresholds for the aberrant chromosomes, with no false negative or positive results. Full-blown fetal aneuploidies can thus be reliably detected in maternal plasma using a multiplex MPS-byligation approach. Furthermore, the results obtained with a sample from a pregnancy with 45,X in the cytotrophoblastic cell layer and 46,XX in the mesenchymal core cells show that ccffDNA originates from the cytotrophoblastic cell layer. Discrepancies between the genetic constitution of this cell layer and the fetus itself are well known, and therefore, care should be taken when translating results to the fetus itself.

INTRODUCTION

For the prenatal detection of chromosomal aberrations, fetal material is needed which, until recently, could only be obtained via invasive procedures such as chorionic villus sampling (CVS) or amniocentesis (AC). These procedures entail a minor but definite risk of miscarriage, and therefore CVS and AC are only offered to pregnant women at relative risk of carrying a fetus with a chromosomal aberration. Since 1997, however, it is known that plasma of pregnant women contains circulating cell-free fetal DNA (ccffDNA) [1]. As this ccffDNA can be obtained via maternal blood sampling, it represents an attractive source of fetal material for non-invasive prenatal diagnosis (NIPD). Within this context, several characteristics of this ccffDNA have been studied extensively. Compared with fetal cells in the maternal circulation, the concentration of ccffDNA in maternal plasma is relatively high and the fetal/maternal DNA concentration ratio favorable (estimates: ~ 3-19% of the DNA in the plasma is of fetal origin [2-4]). This turns ccffDNA into a more attractive source for NIPD than intact fetal cells in the maternal circulation. Additionally, ccffDNA has a

very short half-life time (mean 16.3 minutes) and will, therefore, not be detected in the maternal circulation after delivery [2]. The ccffDNA has been shown to be derived from the placenta and is, therefore, extraembryonic in origin [5-8]. The placental DNA is released into the maternal circulation by apoptosis, explaining the highly specific fragmented nature of ccffDNA [3,9]. Taken together, ccffDNA may offer a range of possibilities for application in NIPD.

In 2008, two research groups independently showed that ccffDNA can be used for NIPD of fetal trisomies using massively parallel sequencing (MPS)-by-synthesis [3,10]. In short, DNA fragments, isolated from maternal plasma samples containing DNA both fetal and maternal in origin, were used for MPS taking advantage of the fact that through MPS several hundreds of millions DNA fragments can be sequenced and quantified at once. These fragments were subsequently mapped to the human reference genome and the (relative) numbers of fragments per chromosome were counted. Hence, if the fetus carries a trisomy, more fragments of the trisomic chromosome are expected to be mapped to the specific chromosome when compared with the other normal (diploid) chromosomes. The initial proof-of-concept studies showed that this method allowed a correct prediction of fetal trisomies 21 (n=23), 13 (n=1) and 18 (n=2), without false negative or positive results [3,10]. Since then, several large-scale studies have shown that MPS-by-synthesis can be used (also in a clinical setting) for non-invasively diagnosing fetal trisomy 21 [11-14]. As such, this test is to date commercially offered to pregnant women in the USA. The efficacy to detect other fetal trisomies, such as trisomy 13 and 18, has also been studied. The detection of these trisomies, however, appeared to be less accurate when the same statistical algorithm as for the detection of trisomy 21 was applied [15], which was attributed to a lower GC content of these chromosomes. However, when using adapted algorithms also trisomies 13 and 18 could be identified correctly with MPS-by-synthesis methods [13,15].

So far, almost all studies dealt with MPS-by-synthesis methods using the Illumina GA(II) or HiSeq 2000 platforms. It is, however, known that sequence performance and ability to multiplex samples depend on the MPS platform used. In our laboratory, MPS-by-ligation is routinely performed using a SOLiD 4 platform (Applied Biosystems/Life Technologies, Foster City, CA, USA) in both postnatal diagnostic and research settings. As such, we aimed in the present study to technically evaluate the possibility to perform NIPD with ccffDNA for the detection of fetal aneuploidies using MPS by-ligation, simultaneously testing the possibility of multiplexing samples. Moreover, we established the placental cell layer origin from which cffDNA is derived.

Table 1. *Overview of the samples from pregnancies with fetal aneuploidies, including run number, gestational ages and Z-scores for the aberrant chromosomes with different algorithms.*

Sample number	Aberration	Run number	Gestational age	Collected before or after an invasive procedure	Aneuploid autosomal Z-scores	
					ALG1	ALG2
A	Trisomy 13	35	13.2	5 days after CVS	9.12	-
B	Trisomy 13	109	19.2	Before	6.25	5.10
C	Trisomy 18	35	15.1	10 days after CVS	11.60	-
D	Trisomy 18	45	13.0	Before	13.62	-
E	Trisomy 18	45	20.3	8 days after AC	9.52	-
F	Trisomy 18	63	12.6	7 days after CVS	4.70	5.14
G	Trisomy 21	18	12.4	6 days after CVS	5.28	-
H	Trisomy 21	35	19.6	7 weeks after CVS	7.29	-
I	Trisomy 21	45	21.1	6 days after AC	3.12	-
J	Trisomy 21	45	15.4	Before	7.51	-
K	Trisomy 21	45	12.4	4 days after CVS	5.80	-
L	Trisomy 21	63	20.3	8 weeks after CVS	5.80	11.81
M	Trisomy 21	63	13.2	2 days after CVS	8.29	16.56
N	Trisomy 21	63	15.5	Before	3.38	7.17
O*	Trisomy 21	109	13.5	Before	4.73	4.62
P	Trisomy 21	109	10.5	Before	10.22	10.13
					X-chromosome Z-scores	
Q	47,XXX	35	19.1	3 weeks after AC	5.23	
R	CVS: 45,X[8] (STC) and 46,XX[25] (LTC) AC: 46,X[1]/ 46,XX[28] Newborn blood: 46,XX[30]	63	16.1	20 days after CVS, before AC	-7.72	
S	45,X	109	11.1	before	-5.62	

** Processing of this sample was carried out 7½ h after collection of the blood.*
AC: Amniocentesis; CVS: Chorionic villus sampling; LTC: Long-term culture; STC: Short-term culture; -: Not determined.

STUDY DESIGN / MATERIALS AND METHODS

SUBJECT ENROLLMENT

Pregnant women were recruited at the Department of Obstetrics and Gynecology, Radboud University Nijmegen Medical Centre, Nijmegen, and the Department of Obstetrics and Gynecology, Rijnstate Ziekenhuis, Arnhem, The Netherlands. The study was approved by the ethical committee. All pregnant women received written and oral information regarding the study and from all pregnant women signed informed consent was obtained.

PATIENTS AND SAMPLES

From 52 pregnant women (singleton pregnancies for which an invasive procedure was indicated) at various gestational ages blood samples (20-40 ml in EDTA (ethylenediaminetetraacetic acid) anti-coagulated tubes) were collected. Genetic fetal analyses were diagnostically carried out on AC or CVS cells by either QF-PCR only (Aneufast kit, version 2, Genomed Ltd., Kent, UK), routine karyotyping (both according to standard protocols) or 250k SNP array analysis (according to the protocol described by Faas et al. [16]). Fetal karyotypes from samples from abnormal pregnancies, always confirmed by routine diagnostic karyotyping, are listed in **Table 1**. These samples were drawn either before or after an invasive procedure. All blood samples from pregnancies with normal fetuses were drawn before an invasive procedure. The mean gestational age of the normal pregnancies was 12.2 weeks (range 8.5-16.1 weeks): 28 samples were from normal first trimester pregnancies, 5 from second trimester pregnancies.

All blood samples were kept on room temperature before processing and all, except for sample O, were processed within 6 h after collection by centrifugation at 1600g for 10 minutes, separating plasma from the buffy coat. Sample O was processed after ~7½ h. The plasma was then transferred to 1.5 ml Eppendorf tubes and recentrifuged at 16,000g. The supernatants were collected and stored at -80 °C until further processing.

DNA ISOLATION

Plasma DNA was isolated from 3.2 to 7 ml of the frozen plasma samples using a QIAamp DSP DNA Blood Mini Kit or a QIAamp Circulating Nucleic Acid Kit (QIAgen, Westburg BV, The Netherlands) with minor modifications (exact protocol available on request). Briefly, each plasma sample was divided into two aliquots and for each of these aliquots one isolation column was used. DNA was eluted in 100 µl of low TE buffer (Applied Biosystems, Carlsbad, CA, USA) in Eppendorf DNA LoBind tubes. Of note, the elute of the first column was used as elution buffer for the corresponding second column. The total amount of isolated DNA varied from 8.9 to more than 100

ng per sample. DNA quality was checked using a Bioanalyzer High Sensitivity assay (Agilent Technologies, Waldbronn, Germany). After isolation, DNA samples were stored at 4C until further use.

MASSIVELY PARALLEL SEQUENCING-BY-LIGATION

Library preparation; Sequencing libraries were prepared individually following the Applied Biosystems protocol for Barcoded Fragment Library preparation (version March 2010), without performing the sonication of DNA for generating short fragments, as the ccffDNA is already fragmented in nature. For testing efficacy of multiplexing, all samples but those in run 18, were labeled using unique identifier tags or barcodes (Agilent) that were compliant with SOLiD sequencing technology.
Bead enrichment; The four libraries sequenced in run 18 were not barcoded and, as such, were individually subjected to a manual emulsion PCR and bead enrichment (SOLiD 4 System Templated Bead Preparation QRC, Rev B 03/2010. Remark: 0.5 pM used as input for emPCR). For multiplexed sequencing runs (runs 35 (8-plex), 45 (16-plex), 63 (16-plex) and 109 (16-plex)), finished libraries were equimolarly pooled for multiplexing up to 16 samples simultaneously with a final combined library concentration of 700 pM. Subsequently, the obtained pool was used for emulsion PCR (E80 scale) and bead preparation using the EZ bead system, following manufacturer's instructions (version May 2010; Life Technologies).

SEQUENCING

Massively parallel whole genome sequencing-by-ligation was performed on a SOLiD 4 System (Life Technologies, Foster City, CA, USA), according to the manufacturer's protocol. After 3' end modification, enriched beads were quantified using a Nanodrop. For each of the four samples in run 18, one quad of a sequencing slide was used, whereas for each multiplexed pool (runs 35, 45, 63 and 109), a full sequencing slide was used (Life Technologies, Carlsbad, CA, USA). For multiplex sequencing runs, we anticipated that all samples in the pool would be equally represented in the total number of beads present on the sequencing slide.

MAPPING

The 50-bp color space reads were mapped to the hg19 reference genome with the SOLiD bioscope software v1.3, which utilizes an iterative mapping approach. Fragments of which at least 30 bp were mapped were included for further analyses. Additionally, only reads with a minimum mapping quality (MAPQ) of 60 (PHRED scaled) were used for read counting (allowing a maximum of 1 mismatch for fragments with read lengths between 30 and 47 bp and a maximum of 2 mismatches for fragments with read lengths between 48 and 50 bp). These quality criteria result

in mapped reads with a 0.001% chance of being misaligned, thereby drastically reducing multiple mapping artifacts.

DATA ANALYSIS

AUTOSOMAL CHROMOSOMES

For all samples, Z-scores were calculated as previously described [10]. First, the percentage of reads of a chromosome of interest of the total number of reads of a sample was determined, using the following equation:

$$\%a = \frac{\#\text{ reads on a}}{\#\text{ reads on all chromosomes}} * 100$$

a = chromosome of interest
# reads = reads passing the MAPQ filter step (within a single sample)

Subsequently, this percentage was compared with the mean percentage of the same chromosome of a set of reference samples, and Z-scores were calculated using the following equation:

$$\text{Z-score} = \frac{\%a - \text{mean (reference }\%a\text{)}}{\text{SD (reference }\%a\text{)}}$$

a = chromosome of interest
reference %a = values as observed in control samples

For algorithm 1 (ALG1), all samples in a run, regardless of the fetal karyotype, were used as reference samples. As this also includes samples from pregnancies with abnormal karyotypes, for the calculation of the mean of a specific chromosome, the aberrant sample was excluded. For algorithm 2 (ALG2), mean percentages were calculated by combining results from normal males from different runs as references. Thresholds for aneuploidy were set at $Z > +3$ and $Z < -3$ for over- or under representation, respectively.

DETERMINATION OF FETAL GENDER

Similar to others [10,11,17], we also detected the presence of Y-chromosome-specific sequences in samples from female pregnancies. Therefore, for determination of the fetal gender, first the percentages of reads of both the X- and Y-chromosomes

in reference males and females were determined as described above (see Section 2.5.1). Obtained values were used to determine the indicative value for the presence of a Y-chromosome and, subsequently, fetal gender of the samples was predicted. For all samples, determined as not carrying Y-chromosome specific sequences, the X-chromosome Z-score was calculated, using an algorithm comparable with ALG1, with normal females as references (in case insufficient samples from normal female fetuses were available in one run, samples from (autosomal) chromosomally abnormal fetuses were also used as a reference). This algorithm was termed ALGX.

RESULTS

TECHNICAL FACTS

In this study, samples sequenced in 4-, 8- and 16-plex runs were included. Runs 18, 35 and 45 were considered to be runs in the learning phase. Runs 63 and 109 were technically considered optimal and in these runs the mean number of good quality reads per sample that met our criteria was 18.3 x 10^6 (standard deviation (SD) 3.8 x 10^6: maximum 23.5 x 10^6, minimum 9.5 x 10^6; on average 63% of the number of raw reads).

DETERMINATION OF AUTOSOMAL (AN)EUPLOIDIES

ALG1 was applied to all runs and for all samples and all autosomal chromosomes revealed Z-scores as expected: all trisomies 13 (n=2), 18 (n=4) and 21 (n=10) showed Z-scores > 3.0 for the trisomic chromosomes with Z-scores within the normal range for the other chromosomes (**Table 1**). ALG1, however, requires pre-test knowledge of the fetal karyotypes, as for calculation of the mean percentage of reads of a chromosome of interest all samples in a run are included and in case of an aberration, an adjustment for the aberrant chromosome is carried out. Therefore, for runs 63 and 109 Z-scores were also calculated with ALG2, for which no pre-test knowledge is necessary. This yielded Z-scores for all samples as expected (**Table 1 & Figure 1**), with no false positive or negative results. The standard deviation of the means of the various chromosomes in the reference male samples, indicative for the variation between normal samples, ranged from 0.006 for chromosome 14 to 0.210 for chromosome 4. For chromosomes 13, 18 and 21, this was 0.084, 0.025 and 0.008, respectively. **Figure 2** shows a schematic overview of the standard deviations of the means using ALG2.

DETERMINATION OF SEX CHROMOSOMES

For runs 18, 35 and 45, the percentages of Y-chromosome specific sequences varied per run. Nevertheless, based on the percentages in normal males and females,

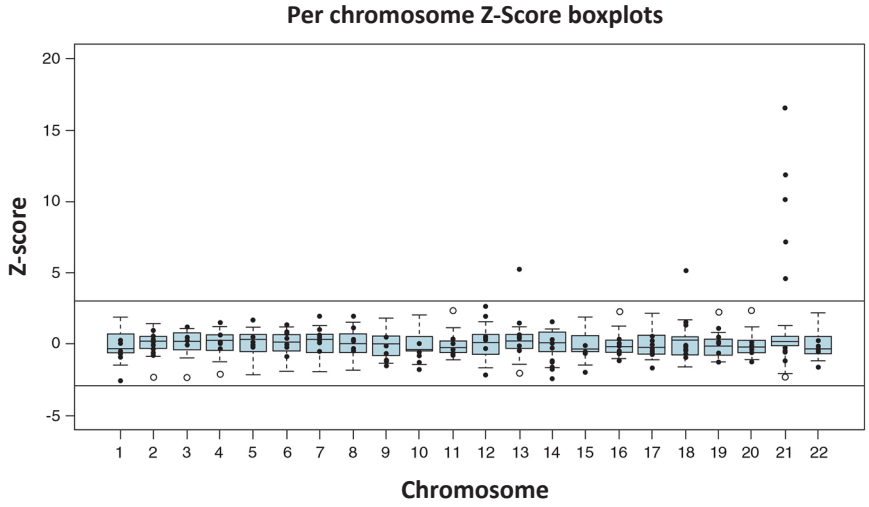


Figure 1. *Boxplot with Z-scores of the autosomal chromosomes, when applying ALG2. The samples with Z-scores > 3 are the samples from runs 63 and 109 with abnormal fetal karyotypes, as listed in Table 1.*

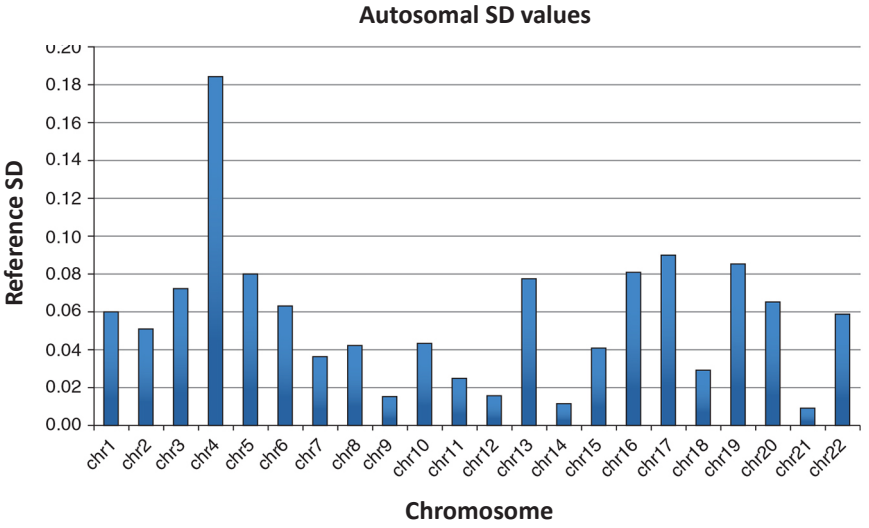


Figure 2. *Standard deviations (SD) of the means of the autosomal chromosomes, when applying ALG2.*

for all three runs a relatively high or low percentage of Y-chromosome-specific sequences could be correctly correlated to the presence or absence, respectively, of a Y-chromosome. Both samples Q (fetal 47,XXX; run 35) and R (fetal 45,X/46,XX; run 45 (also included in run 63)) showed less Y-specific sequences than samples from normal male pregnancies, indicating the absence of a Y-chromosome. For runs 63 and 109, the mean percentages of Y-chromosome specific reads in samples from normal fetal males and females were comparable in both runs and combining both runs resulted in 0.018% (SD 0.001: highest 0.019%, lowest 0.016%) for samples from normal fetal females and 0.033% (SD 0.009: highest 0.052%, lowest 0.025%) in samples from normal fetal males. The presence or absence of a Y-chromosome could be predicted correctly in all samples and both sample R (fetal 45,X/46,XX; run 63) and S (fetal 45,X; run 109) revealed percentages of Y-chromosome specific sequences comparable with normal female samples (0.023 and 0.019%, respectively). Subsequently, for samples predicted as not having Y-chromosome-specific sequences, Z-scores for the X-chromosome were calculated, according to ALGX. The fetal sex was predicted correctly in all cases, including samples Q (fetal 47,XXX: X-chromosome Z-score 5.27), R (fetal 45, X/ 46,XX: X-chromosome Z-score -7.73) and S (fetal 45,X: X-chromosome Z-score -5.62).

CONCLUSIONS

From this technical evaluation study we conclude that for the non-invasive detection of fetal autosomal aneuploidies from ccffDNA in maternal plasma MPS-by-ligation with the SOLiD 4 platform can be applied, and that sequencing of up to at least 16 samples simultaneously is possible on this platform. As the first runs (runs 18, 35 and 45) contained relatively few samples from normal pregnancies that could serve as reference samples, ALG1 was applied. With this algorithm the presence or absence of autosomal aneuploidies could be predicted correctly. ALG1, however, requires a priori knowledge of the fetal karyotypes and thus cannot be used in a clinical diagnostic setting. Therefore, after this learning phase, another algorithm for which no pre-test knowledge is necessary (ALG2), was applied. With this algorithm, only samples from normal male pregnancies were used as references to calculate the mean percentage of reads of a specific chromosome and, subsequently, Z-scores from all samples from two different runs were calculated: all fetal autosomal aneuploidies, including trisomies 13, 18 and 21, could be predicted correctly with no false positive or negative results. Also, the fetal sex and sex chromosomal aneuploidies could be predicted correctly by first determining the presence or absence of a Y-chromosome and, subsequently, determining the number of X-chromosomes in the Y-chromosome-negative samples.

Additionally, evidence is provided that ccffDNA in the maternal plasma is derived from the cytotrophoblastic cell layer of the placenta, as results obtained with sample R are suggestive for the presence of one X-chromosome in the absence of a Y-chromosome. This is in line with the fetal karyotyping results obtained with cells from the cytotrophoblastic cell layer of the placenta, and not of the mesenchymal core cells of the placenta, neither of the AC cells nor the cells from the newborn blood.

EXPERT OPINION

Cell-free fetal DNA is present in the plasma of pregnant women in relatively large amounts and in a relatively favorable percentage of the total amount of DNA in the plasma. As this DNA can be obtained via non-invasive procedures, it represents an attractive source for NIPD. So far, several studies have shown the possibility of using MPS-by-synthesis with cffDNA for the detection of fetal trisomy 21, including two large-scale studies in a clinical setting [3,10-14]. Since not all laboratories are equipped with the same platform and different platforms may exhibit different performances, we set out to test the efficacy of the Applied Biosystems SOLiD 4 platform. To our best knowledge, so far only one study has described the successful use of MPS-by-ligation for NIPD of trisomy 21 using the SOLiD 3 system of Applied Biosystems [17]. In the present study, we confirm their results with MPS-by-ligation, in our setting using the SOLiD 4 platform, and show on 52 samples that this platform indeed can be applied for the robust and reproducible detection of fetal aneuploidies with ccffDNA from plasma of pregnant women.

As can be seen from **Table 1**, not all aneuploid samples were collected before an invasive procedure. One might argue this influences the results, as additional fetal DNA might be released into the maternal circulation as a consequence of the invasive procedure. To the best of our knowledge, however, there exists no evidence for such a procedure related increase in ccffDNA. Moreover, the blood was always collected several days up to several weeks after CVS or AC. Even if there had been a procedure-related increase of ccffDNA, this would not be detectable anymore in our samples, because of the very short half-life time of the cffDNA [2].

The present study was not restricted to the detection of fetal trisomy 21, as two fetal trisomy 13 and four fetal trisomy 18 cases were included and diagnosed correctly too, using the same algorithm as for the diagnosis of fetal trisomy 21. Previously, others argued that this algorithm might be less accurate for diagnosing fetal trisomy 13 and 18 [15] as, due to the GC content of these chromosomes, there is a broader sample to sample variation in number of reads per chromosome, as compared with chromosome 21. Therefore, they suggested the use of an algorithm with GC-correction for calculating Z-scores [15]. Also Sehnert et al. used a different

approach to correctly identify fetal trisomy 18 [13]. In the present study, we indeed observed a broader sample-to-sample variation for chromosomes 13 and 18 than for chromosome 21, as reflected by the SD of the means, which were 0.084, 0.025 and 0.008, respectively (**Figure 2**). We were, however, able to correctly predict fetal aneuploidies in all cases using ALG1 or ALG2, suggesting that with MPS-by-ligation using the SOLiD 4 platform, no GC-correction is necessary for reliable diagnosis of fetal trisomy 13, 18 or 21. We do, however, realize that our conclusion might be biased by our relatively small sample size for trisomies 13 and 18 (n=2 and 4, respectively).

In our setting, we multiplexed up to 16 samples, which resulted in ~ 18 x 10^6 good quality reads per sample, with ~ 9 x 10^6 reads being the lowest number of reads achieved per sample. Previously, others showed with MPS by-synthesis that a minimum of between 0.3 and 2.3 x 10^6 good quality reads per sample are necessary for reliable results [11]. As we expect this to be comparable for MPS-byligation it must, theoretically, be possible to obtain a threefold higher multiplexing sequencing rate than presented here, thereby significantly reducing costs. Further studies to test this are in motion. Even though the results of the present study are in full concordance with the results obtained with invasive prenatal diagnosis, we did not include a test to check for the presence of fetal DNA, in case a normal female fetus was predicted. Before the test is implemented in routine daily prenatal practice, such a check, as also described by others [12], should be incorporated.

In summary, we conclude that the issue is no longer whether NIPD for fetal aneuploidies is possible, as we and others have now shown that both MPS-by-synthesis and MPS-by-ligation can be reliably used, but how it can be implemented in the daily practice of prenatal diagnosis. Nowadays, the first trimester screening is offered routinely to pregnant women. With this non-invasive test, however, a risk on fetal trisomy 13, 18 and 21 is calculated, whereas NIPD by MPS results in definite diagnosis. Even though for the time being first trimester screening is a much cheaper and quicker method than NIPD with MPS, the latter is expected to replace the first trimester screening in the near future, as prices will drop and turnaround times of MPS will improve. As a result of the much higher predictive value of NIPD with MPS, the number of invasive procedures will probably be reduced significantly as invasive procedures will mostly be restricted to those pregnant women with a positive NIPD result. Nuchal translucency (NT) measurement is currently used in first trimester screening as part of risk assessment for trisomy 13, 18 and 21, but this measurement alone may also identify pregnancies at a high risk of non-chromosomal disorders, such as congenital heart defects and monogenetic disorders, such as Noonan syndrome [18]. Therefore, one could opt for keeping NT measurement in place as a separate screening test, complementary to NIPD.

Even though ccffDNA in the maternal plasma is a very attractive source of fetal DNA for NIPD, one should be aware of the fact that the fetal DNA that is studied with NIPD is extraembryonic in origin and reflects the genetic constitution of the cytotrophoblastic cell layer of the placenta, rather than the mesenchymal core cell layer. Discrepancies between the cytotrophoblastic cell layer of the placenta and the fetus itself are well-known. Therefore, one might opt for a strategy in which all NIPD-positive cases are offered AC, or at least those cases of aneuploidy of which discrepancies between the cytotrophoblastic cell layer and the fetus itself are known to occur (e.g., trisomy 18).

ACKNOWLEDGEMENTS

The authors thank MFWJ Ariaans, CA van Erp and NJA Diependaal, M End-van Arkelen and C Jurrius for recruiting pregnant women, C Keesmaat, J Willemen-Derks and I Gomes as well as all other technicians of the Prenatal Diagnostics Group of the Department of Human Genetics, Radboud University Nijmegen Medical Centre, and dr PMW Janssen from the Laboratory for Clinical Chemistry, Rijnstate Ziekenhuis, Arnhem, The Netherlands, for their support and dr. ir. JA Veltman for his expert advice and support. This work was in part supported by grants from the Netherlands Organization for Health Research and Development (ZonMW 916.86.016 to LV) and the EU funded TECHGENE project (Health-F5-2009-223143 to JdL).

DECLARATION OF INTEREST

The authors state no conflict of interest and have received no payment in preparation of this manuscript.

REFERENCES

1. Lo YM, Corbetta N, Chamberlain PF, et al. Presence of fetal DNA in maternal plasma and serum. Lancet 1997;350:485-7
2. Lo YM, Tein MS, Lau TK, et al. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. Am J Hum Genet 1998;62:768-75
3. Fan HC, Blumenfeld YJ, Chitkara U, et al. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA 2008;105(42):16266-71
4. Lun FM, Chiu RWK, Chan KCA, et al. Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. Clin Chem 2008;54:1664-72
5. Ng EK, Tsui NB, Lau TK, et al. mRNA of placental origin is readily detectable in maternal plasma. Proc Natl Acad Sci USA 2003;100:4748-53
6. Jackson L. Fetal cells and DNA in maternal blood. Prenat Diagn 2003;23:837-46

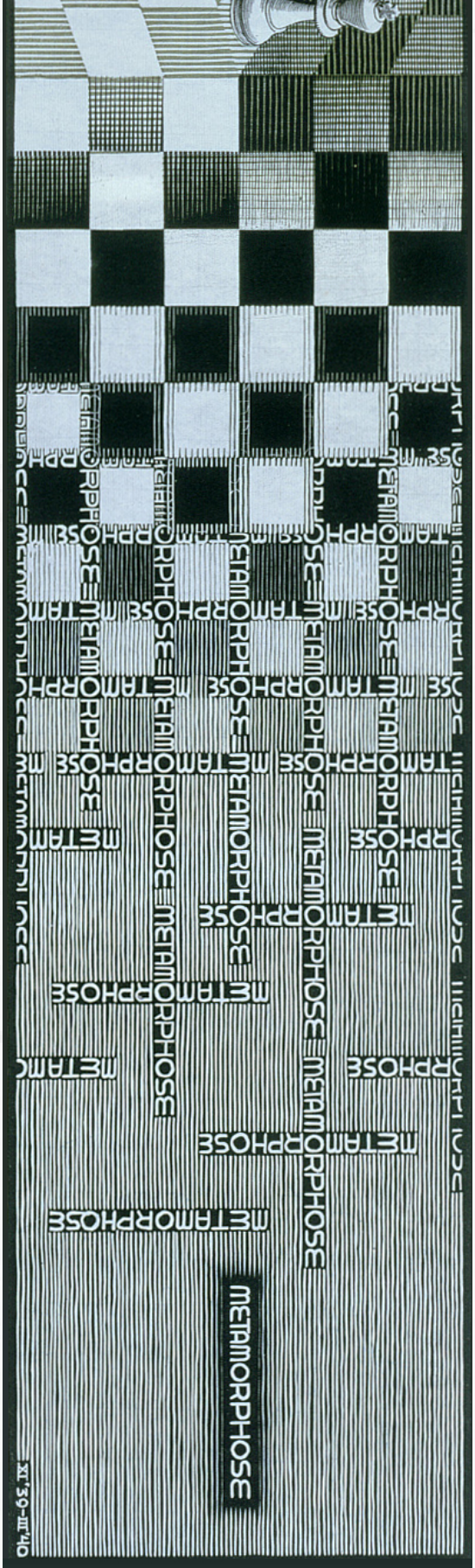
7. Masuzaki H, Miura K, Yoshiura K-I, et al. Detection of cell free placental DNA in maternal plasma: direct evidence from three cases of confined placental mosaicism. J Med Genet 2004;41:289-92
8. Alberry M, Maddocks D, Jones M, et al. Free fetal DNA in maternal plasma in anembryonic pregnancies: confirmation that the origin is the trophoblast. Prenat Diagn 2007;27(5):415-18
9. Lo YM, Chan KC, Sun H, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med 2010;2(61):61ra91
10. Chiu RWK, Chan KCA, Gao Y, et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci USA 2008;105(51):20458-63
11. Chiu RWK, Akelokar R, Zheng YWL, et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validation study. BMJ 2011;342:c7401
12. Ehrich M, Deciu C, Zwiefelhofer T, et al. Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. Am J Obstet Gynecol 2011;204:205.e1-11
13. Sehnert AJ, Rhees B, Comstock D, et al. Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood. Clin Chem 2011;57(7):1042-9
14. Palomaki GE, Kloza EM, Lambert-Messerlian GM, et al. DNA sequencing of maternal plasma to detect Down syndrome: an international clinical validation study. Genet Med 2011;13(11):913-20
15. Chen EZ, Chiu RWK, Sun H, et al. Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. PLoS ONE 2011;6(7):e21791
16. Faas BHW, van der Burgt I, Kooper AJ, et al. Identification of clinically significant, submicroscopic chromosome alterations and UPD in fetuses with ultrasound anomalies using genome-wide 250k SNP array analysis. J Med Genet 2010;47(9):586-94
17. Chiu RWK, Sun H, Akolekar R, et al. Maternal plasma DNA analysis with massively parallel sequencing by ligation for noninvasive prenatal diagnosis of trisomy 21. Clin Chem 2010;56(3):459-63
18. Bilardo CM, Timmerman E, Pajkrt E, et al. Increased nuchal translucency in euploid fetuses--what should we be telling the parents? Prenat Diagn 2010;30(2):93-102

AFFILIATIONS

Brigitte HW Faas†[1,6] PhD, Joep de Ligt[1,7] MSc, Irene Janssen[1,8], Alex J Eggink[2,3,9] MD PHD, Lia DE Wijnberger[5,10] MD PhD, John MG van Vugt[2,4,11] MD PhD, Lisenka Vissers[1,12] PhD & Ad Geurts van Kessel[1,13] PhD

†Author for correspondence

1. Radboud University Nijmegen Medical Centre, Department of Human Genetics, PO Box 9101, 6500 HB Nijmegen, The Netherlands Tel: +024 3614104; Fax: +024 3668751; E-mail: b.faas@gen.umcn.nl
2. Radboud University Nijmegen Medical Center, Department of Obstetrics and Gynecology,
3. Nijmegen, The Netherlands
4. University Medical Centre Rotterdam, Department of Obstetrics and Gynecology, Erasmus MC, The Netherlands
5. Also on behalf of the gynecologists of the Network for Prenatal Diagnosis Nijmegen, Nijmegen, The Netherlands
6. Rijnstate Hospital, Department of Gynecology and Obstetrics, Arnhem, The Netherlands
7. Clinical cytogeneticist, Radboud University Nijmegen Medical Centre, Department of Human Genetics, Nijmegen, The Netherlands
8. Bioinformatician, Radboud University Nijmegen Medical Centre, Department of Human Genetics, Nijmegen, The Netherlands
9. Research technician, Radboud University Nijmegen Medical Centre, Department of Human Genetics, Nijmegen, The Netherlands
10. Perinatologist-gynaecologist, Radboud University Nijmegen Medical Center, Department of Obstetrics and Gynecology, Nijmegen, The Netherlands
11. Perinatologist-gynaecologist, Rijnstate Hospital, Department of Gynecology and Obstetrics, Arnhem, The Netherlands
12. Professor, Perinatologist-gynaecologist, Radboud University Nijmegen Medical Center, Department of Obstetrics and Gynecology, Nijmegen, The Netherlands
13. Postdoc, Radboud University Nijmegen Medical Centre, Department of Human Genetics, Nijmegen, The Netherlands
14. Professor, Clinical cytogeneticist, Radboud University Nijmegen Medical Centre, Department of Human Genetics, Nijmegen, The Netherlands



Artwork reproduced with permission from the M.C. Escher Company.

Chapter 6

Discussion

THE EVOLUTION OF GENETICS AND (BIO)INFORMATICS

In 1860 the study of genetics began with Gregor J Mendel investigating the inheritance of traits in pea plants [1]. As time progressed, geneticists were able to study more traits with, for example Drosophila, studied in the famous fly room of Thomas H Morgan [2]. Morgan envisioned traits as being physically linked and this resulted in the creation of the first chromosome map in 1911 by Alfred Sturtevant [3]. The concept of observing and mathematically describing traits is of particular relevance for the application of computational methods to the genetics field [2]. Studying inheritance in plants and animals was essentially of a quantitative nature, therefore many early geneticists used and developed statistical and quantitative methods for their research [2].

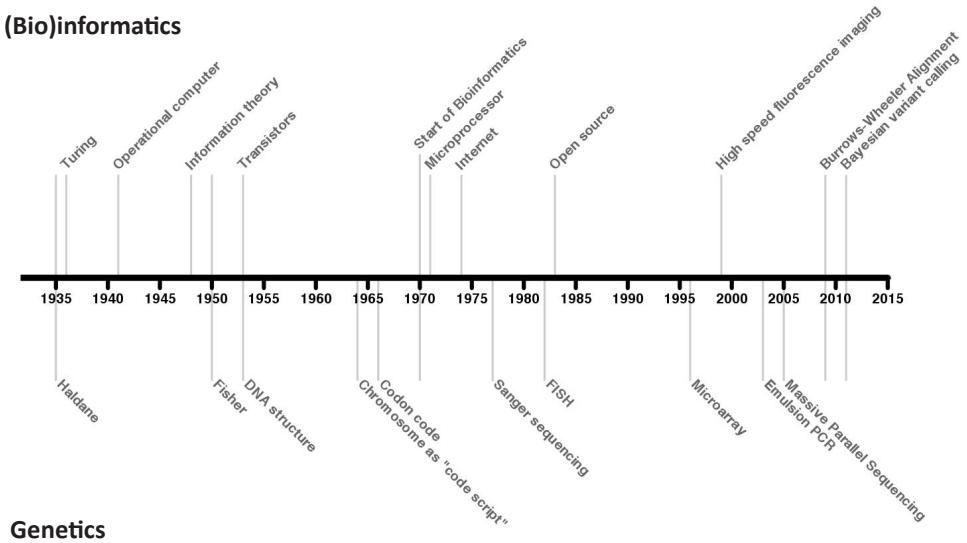


***Figure 1**, timeline depicting important breakthroughs for the fields of (bio)informatics (top) and genetics (bottom).*

Amongst the many mathematicians who brought statistics to the genetics field [2], there is one of special interest for the emergence of bioinformatics; John J Haldane, famous for his early estimates of human mutation rates [4] (**Figure 1**). Haldane's observations demonstrated the power of applying mathematical models to biological observations. One other important mathematician is Alan Turing, one of the founding fathers of information theory who formalized the concepts of algorithms and computation [5]. Turing provided the means to apply these statistics to more complex problems in a structured and systematic manner, one of the foundations of bioinformatics. The early successes of statistics led to the

Box 1, Schrödinger's cat refers to a thought experiment which illustrates how probabilities resolve into fixed states [1], it describes a scenario in which an imaginary cat is equally likely to be dead or alive (in terms of probabilities), while in reality one immediately realizes that the cat must be either one or the other. The original experiment served to illustrate a paradox in quantum mechanics, it can now be applied to genetic testing results. For example; when considering the risk of disease, an individual can have equal risks of being affected or not while in reality that person will be either affected or healthy. We can use probabilities to predict something about a group of individuals well, while it can be difficult to apply to a single individual. Until all factors contributing to risks and modifiers are understood interpreting chance will be an essential part of genetic testing and counseling.

[1] Schrödinger E. Die gegenwärtige Situation in der Quantenmechanik Naturwissenschaften 23, 807-849 (1935). (The present situation in quantum mechanics). Translation available at; www.tuhh.de/rzt/rzt/it/QM/cat.html

merging of genetics and complex mathematics, ultimately resulting in the field of bioinformatics. Bioinformatics uses (bio) statistics to test hypotheses on large quantities of data, through a large number of complex calculations. The field distinguishes from biostatistics by the incorporation of informatics concepts such as neural networks, graph theory, process optimization, file handling, database operations, prediction models and detection algorithms [2,6].

Similar to the concept of traits having a physical distance, the concept of DNA as a code formed an important link between genetics and information theory. In 1946 Erwin Schrödinger (owner of a well known imaginary cat [**box 1**]) envisioned chromosomes to contain a "code-script" [7]. According to Schrödinger's reasoning an entity which could read and interpret the chromosomal code would be able to "tell from their structure whether the egg would develop … into a black cock or a speckled hen …" [2,7]. This notion of DNA code is highly similar to the concepts introduced by Turing in his work on algorithms; a structured set of commands to perform complex calculations encoded in a computer [5].

Subsequent research of the 'chromosome code' uncovered a high degree of similarity between computer and genetic code; DNA encodes information in four nucleotides; A, T, C and G and computers encode information binary (1 or 0) [2]. While DNA codes are chemically encoded and computer codes mostly electronically, both methods use a relative simple set of elements as building blocks to construct an immense quantity of complex commands and structures.

The first application of a programmed computer to solve a biological problem was published in 1950 [8] by Ronald A Fisher (therefore arguably the first bioinformatician), in which he used differential equations to compute gene frequencies in populations. Many of Fisher's statistical tests, like the Fisher exact test and his analysis of variance (ANOVA) models are still used in genetic research to test hypotheses [2,6,9].

Both bioinformatic and genetic research accelerated quickly through the work of James Watson and Francis Crick, who described the structure of DNA in 1953 [10]. The structure of DNA was described based on experimental data from Rosalind Franklin and Raymond Gosling [11], supported by the structure observed by Maurice Wilkings and colleagues [11]. In these early days, bioinformatics was perceived by many as a tool to process large quantities of data [12]. Biomedical researchers involved bioinformaticians after a hypothesis was formed and data were gathered to calculate quantitative values, for example, on sequence similarity [2,6]. The role of bioinformatics soon evolved into a more exploratory one by studying how DNA encoded proteins and to identify where these coding elements (genes) resided in the genome [6]. Through combined genetics and informatics efforts, the definition of a gene changed from an abstract unit of inheritance to a set of structured DNA sequences [13].

Similar to computer code, DNA sequences were found to contain operators telling the cells translation machinery where to start and stop translation [13]. However, it remained unclear how these DNA sequences encoded amino acids resulting in protein sequences. The presence of codons, encoded by a triplet sequence, was not immediately evident from the initial DNA sequences and was not solved in spite of many years worth of computational modeling [14]. Biologists in the lab of Marshall W Nirenberg eventually identified the first nucleotide triplet code for amino acid encoding in 1965 [15]. The identification of all codons by 1966 is referred to as “the cracking of the genetic code” [15], enabling researchers to start reading the protein coding parts of DNA [15]. The fields of genetics and bioinformatics were now collaborating more closely and by 1970 the first algorithms to perform sequence alignments were published [16,17]. The great interest in DNA sequences led to the introduction of the revolutionary sequencing technique developed by Frederick Sanger and colleagues in 1977 [18].

Computational methods were also spreading to other biology disciplines such as evolutionary biology using for example phylogenetic trees, proteomics using protein structure refinements and cellular biology using pathway analysis [6]. The high degree of integration of bioinformatics and genetics resulted in a blurring of the underlying disciplines [19], leading to changes in the perception of the bioinformatics field.

In 2003, Ouzounis and Valencia reviewed the rise and future of bioinformatics and stated “this discipline will continue to evolve rapidly into the 21st century, perhaps to a point beyond recognition” [6]. Indeed the field of bioinformatics has changed rapidly [19], geneticists now see computational analyses as “dry-lab” experiments, complementing and equal partner to “wet-lab” experiments [20,21]. I envision this trend to continue, and as a consequence every scientist in genetics will need to

understand the fundamental principles of information theory, algorithms, databases and statistical models.

While Sanger sequencing was well suited to study single genes it was not until new genetic technologies became available that researchers were able to routinely take a more holistic approach; Rather than studying one or a few genes they could now study all genes in a certain pathway or in an entire genome [2,22]. These large-scale approaches in genetics are referred to as genomics, the cousin of many other holistic approaches used in biology, collectively called 'omics' approaches [2]. The use of genomics approaches such as microarrays, aCGH and MPS, revolutionized genetic studies and caused an exponential increase in data and need for bioinformatics [19]. In first instance, these techniques required bioinformatics to handle the raw data coming from the genomics instruments. In such large-scale experiments error-correction and noise-reduction models are essential to discern between signal and noise. For example models including a correction for GC content provided more accurate signals for microarray experiments while Hidden Markov Models enabled automated CNV identification in aCGH data [23,24]. In MPS experiments, read correction models and probability based variant identification methods were crucial for the reliable detection of genomic variants [25,26]. With this important role in the early stages of the experiment, bioinformatics has become an integral part of genetics research. A consequence of this is that improvements in data analysis have a substantial impact on the correct identification and subsequent interpretation of genomics data [25].

While it is safe to predict that the role of bioinformatics in genetics will continue to evolve, it is hard to predict exactly how or when. Predictions tied to technological advances are notorious for their limited validity. Such an example is the statement issued by Thomas Watson, president of IBM about computer usage in 1943: "I think there is a world market for maybe five computers". This estimate was perfectly logical at the time when computers were as large as a room, now however it seems foolish in the view of a market size of over one billion computers worldwide. For this statement the large discrepancy between prediction and reality is attributed to the rapid decrease in computer size coupled to an exponential increase in computational power [27].

It is interesting to consider how miniaturization in both sequencing and computational approaches will affect genetics and bioinformatics. Genetics can benefit from informatics, as seen from the development of the IonTorrent™ (Life Technologies) technology [28]. This technology directly translates nucleotides, chemically encoded information, into digital information on a semiconductor chip through the measurement of Hydrogen release upon nucleotide incorporation [28]. Interestingly similar developments are also moving in the opposite direction, for

example the use of the DNA sequences to store data efficiently. DNA-based storage uses the four nucleotide encoding which is twice as efficient in terms of data density compared to the traditional binary encoding [29].

BIOINFORMATICS AND THE DETECTION OF GENOMIC VARIATION

Bioinformatics has played a crucial role in the implementation of both aCGH and MPS approaches in human genetic disease research and subsequent applications in diagnostics [25,30-33]. Bioinformatics was crucial for the optimization of CNV detection by aCGH in the beginning of this millennium [34], similar to its role now for CNV detection in MPS data [**chapter 4,**35]. Bioinformatics has also played a major role in the implementation of MPS technology in general, most notably in the following steps; i) sequence read processing ii) sequence read alignment iii) (small) variant identification and iv) variant annotation.

The first step in which bioinformatics played an important role was the sequence read processing, which is highly platform dependent due to the strong influences from the technical set-up including sequencing chemistry [36]. However, as sequencing technologies are becoming more sensitive and robust, with less sequencing errors and integrated error correction steps, post-experiment read-processing steps will become absolute [37]. Secondary steps such as read mapping and variant identification steps rely heavily on the reference genome, since the currently used fragmented short reads (50-175 bps) are not suited for the *de novo* (reference free) reconstruction of a genome.

The reason why fragmented reads are not suited for *de novo* assembly is that assembly requires overlapping unique reads to construct scaffolds and unique scaffold spanning reads to merge these, which in general requires longer reads or specialized libraries [38]. It was recently shown that *de novo* assembly from short-read paired-end genome sequencing misses approximately 16% of the genome, including over 2,300 coding exons [38]. The regions in which both reference based mapping and assembly based methods perform worst are the duplicated sequences [38]. Almost all duplicated sequences (99%) are missing or misaligned due to the inability of short-reads to be unique within these regions due to high degrees of sequence similarity [38].

Efforts continue to be made to delineate duplicated sequences within the human genome and thereby enable researchers to study the individual genomic locations using so called singly unique nucleotides (SUNs) in each of the locations [39,40]. The principle behind SUNs relies on the occurrence of random mutation, as can been exemplified using *SRGAP2*, an important gene in human brain development [41]. At least two duplications of this gene occurred specifically during human development approximately 3.4 and 2.4 million years ago [42]. Even though the copied segments

of the ancestral *SRGAP2* gene are highly (>99.9%) identical, 16 nucleotides were found to have been affected by random mutation, making those regions unique to one of the copies [42].

Through intensive sequencing and genotyping efforts, scientists identified and localized three individual copies of *SRGAP2* in the human genome; the ancestral *SRGAP2* and two derivative copies, *SRGAP2B* and *SRGAP2C* [42,43]. Functional studies showed that the incomplete copies (B and C) are expressed and translated, and result in truncated forms of the original protein [43]. The protein derived from *SRGAP2C* was found to inhibit the ancestral *SRGAP2* protein. This inhibition delays spine maturation and results in increased density and length of neuronal spines, neuronal features which are considered to be human specific [43]. These and other studies have shown the importance of the correct identification of duplicated sequences and their pivotal role in human evolution and genetic disorders [42-45]. It is important to note that the current human genome reference sequence does not account for most of these duplicated sequences, making these inaccessible to current genetic testing [38].

Technological advances will result in sequencing platforms which produce ever longer and more reliable sequencing reads [37,46]. Longer reads will reduce the time and complexity of the mapping process exponentially as they have a higher likelihood to be unique and therefore easier to map. Additionally the availability of longer reads means that fewer reads will be needed to achieve the desired coverage. When reads can be mapped with greater certainty, the variant identification step will also become more accurate as the difference between a technical error (sequencing or mapping induced) and a true variant will be more distinct and the reference bias introduced by mapping will decrease [47]. Both mapping and *de novo* assembly based variant detection will be highly reliable for simple forms of genetic variation such as single nucleotide variants (SNVs). More complex genetic events will, however, remain more challenging [40,48]. The detection and phasing of small (1-1000 bp) InDel events is still being improved [49], while comprehensive SVs is under constant development [48,50].

Per base sequencing costs and error rates will continue to drop and sequence read lengths increase [37,46]. Decreased costs will enable researchers to combine different approaches and technologies, for example short and long insert size libraries, to obtain a more accurate and complete assessment of an individual's genetic variation [51]. The combination of approaches will allow the detection of different types of variants simultaneously and provide valuable data for accurate effect predictions. When sequencing reaches error rates below one error per million bases, and read lengths of several kbs, *de novo* assembly will be feasible and become a realistic alternative to mapping onto the reference genome.

Switching to *de novo* assembly computational time will increase only slightly due the availability of long unique reads, which can span repeat regions and easily merge scaffolds. Additionally long reads can directly provide allelic (phasing) information, which enables more accurate variant detection and interpretation. The greatest advantage of *de novo* assembly-based methods is that read mapping becomes independent of the human reference genome, which enables the reconstruction of rare genetic architectures, which can be population or even patient specific. While genome reconstruction by *de novo* assembly has already seen great improvements [52,53], there will be future challenges for bioinformatics to perform comprehensive variant identification.

Despite the advantages of assembly-based methods, these still rely on a reference or control genome for variant identification, as do reference mapping based methods. The dependency on a reference means that variant identification, and thereby genetic testing, will only be as reliable and complete as the reference genome(s). The current reference genome does not account for many complex forms or variability, such as inversion haplotypes or repeat expansions, present in the human genome [54]. Efforts to improve the reference genome will continue to have an impact on genetic testing results. I expect that more accurate sequencing and mapping combined with ongoing improvements of the human reference genome, through both computational and technical methods, will result in 99% of an individual's genome to be routinely and affordably accessible for sequence comparison within five years.

Thus far, I have considered experiments that sequence DNA derived from a population of cells, such as whole blood. It is, however, known that genetic mosaicism occurs frequently in humans and have great impact on a person's phenotype [55-57]. These forms of genetic variation are more difficult to detect than germline variants due to the signal intensity being determined by the level of mosaicism. Researchers need to test different tissues with high sensitivity to obtain an understanding of when and where the variant occurred [55,58]. New techniques such as single cell sequencing will be instrumental to obtain tissue and cell specific overviews of genetic variation [59]. A recent study using single cell data showed that mosaic CNVs of one Mb or larger affect 13-41% of neuronal cells in a healthy individual, indicating that the effects of mosaicism are especially relevant for human brain development [60].

INTERPRETATION OF GENETIC VARIATION; THE NEXT CHALLENGE FOR BIOINFORMATICS IN GENETICS

It may be expected that sequencing and variant identification will become trivial in the next few years. Once an unambiguous assessment of the genetic variation in an individual's genome can be obtained with a single test, interpretation of variants to explain or predict a (patient's) phenotype will be the foremost remaining challenge. For bioinformatics, this constitutes a shift of focus from identification to interpretation, involving increasingly complex models to capture the underlying biology. The expected exponential increase in genetic data will also result in a greater demand for data storage, data transfer and standardization of data formats to enable research across and between datasets worldwide.

It will be most practical to start with predicting the effect of every single variant in relation to the patient's phenotype. Once the effect of single variants can be predicted reliably scientists can combine predictions for variants in sets of genes, for example in a particular pathway. These intermediate steps will eventually enable the field to face the challenge of predicting the total effect of all genetic variation in an individual's genome [26,61]. There are eight crucial parts in accurately predicting the phenotypic effect of an individual's genetic variation in which bioinformatics will play an important role:

1. Inheritance: Allelic information for every variant
2. Population variation: Large amounts of phased control genomes
3. Phenotype data: Standardized and quantitative phenotypic data
4. Genome annotation: Accurate descriptions of the functional regions in the genome
5. Functional studies: Standardized and quantitative measurements
6. Prediction tools: More diverse and accurate variant effect prediction
7. Systems biology: Assessment of variant effect across different levels of biology
8. Combining predictions: Combined effect predictions across multiple variants

Figure 2 illustrates where the different parts (1-7) could play a role in the prediction of the phenotypic effect of a single variant. For the remainder of this chapter, I will focus on the practical usages and needs of bioinformatics in variant interpretation, since these bear most directly on the work presented in this thesis.

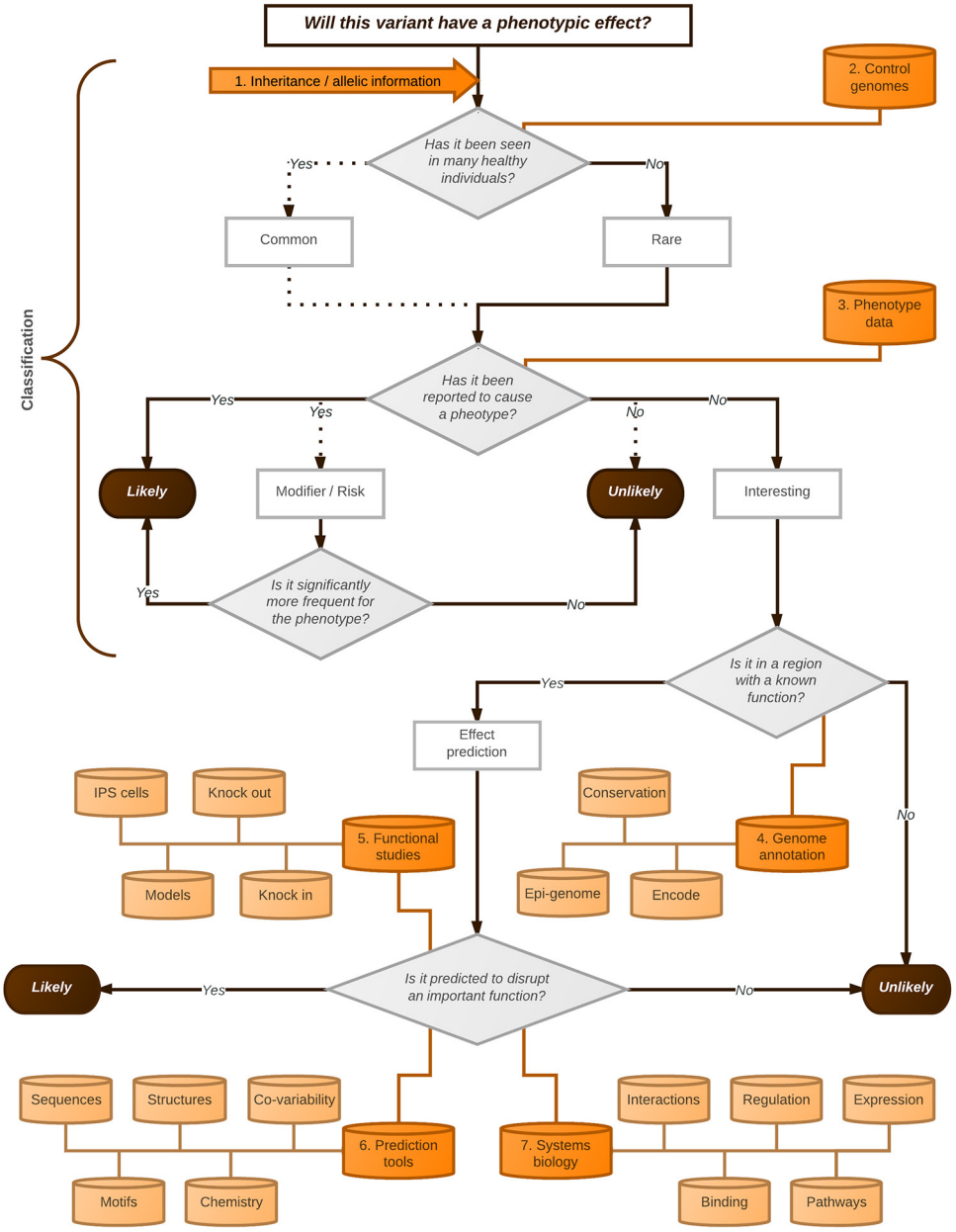


Figure 2. *Conceptual flowchart of a single variant interpretation model. Arrows indicate decisions, yes to the left, no to the right. Dashed arrows indicated the flow for common variants.*

1.) ALLELIC INFORMATION ABOUT EVERY VARIANT

Inheritance information was crucial in this thesis as we tried to explain the sporadic occurrence of intellectual disability (ID) [**chapter 1-5**]. The knowledge of the parental origin or *de novo* occurrence of genetic variants is highly informative in the selection of potential pathogenic variants. Moreover, this information is of utmost importance for the clinic when counseling for recurrence risk as inherited variants will constitute higher recurrence risks compared to *de novo* variants. Additionally, inheritance and allelic information will be an important component for automated variant interpretation models. It is for example clear that two pathogenic variations in a single gene will have a stronger effect on the phenotype when present on different alleles instead of on only one allele.

Bioinformatics can provide allele specific information through the process of phasing, which provides an in silico prediction of the allelic origin. Phasing is starting to be included in variant detection algorithms [36,49], and is more reliable and comprehensive for whole genome data compared to the fragmented data produced by WES and other targeted MPS approaches [62].

The addition of inheritance information to individual and population-based genotype databases will greatly improve their genotyping accuracy and usefulness in interpretation models. Bioinformatic variant calling algorithms can use allelic information to achieve more reliable and complete genotyping of a sample. Inheritance information for large numbers of individuals will also enable bioinformatics to construct models that are more accurate in the assessment of tolerance for genetic variation (mutational load) by enabling the assessment of individual alleles.

2.) LARGE AMOUNTS OF PHASED CONTROL GENOMES

Sequencing healthy individuals will be of great value in variant effect prediction; data on many individuals for a certain position provides a better understanding of the tolerated biological variation at that site. As more and more individual genomes are sequenced, more detailed maps of population structures and their corresponding variant frequencies can be generated. Efforts such as the 1000 genomes project [63], and the Genome of the Netherlands [64] are examples of efforts to establish more accurate datasets cataloguing genome-wide genetic variation. Correcting for factors such as ethnicity and population architecture will be essential for accurate variant effect prediction in a given individual [65].

The effectiveness of control genomes in interpretation is dependent on completeness of the information in the database. Important factors which influence a database's utility are the level of validation of the variants, the level of detail of variant reporting (for example allelic information and ethnicity) and the size of the control dataset.

Information about the validation of a certain variant is crucial in interpretation models. A variant that has been validated (by an additional experiment) and does not have phenotypic effects in a functional assay will be far more informative compared to a variant identified in low coverage MPS data without additional follow-up. Current population frequency databases lack this information as well as inheritance information, limiting their use for variant interpretation.

Studies which provide a complete set of variants per individual, like the personal genome project [66], are of great value for interpretation of the effect of combinations of genetic variants, but they often lack allelic information. Variant co-occurrence information can, for example, be incorporated into so-called mutational load models [67] (see prediction programs section).

Several hundred variants in an individual genome are rare in the general population and therefore a substantial number of control individuals are needed to interpret such rare variants based on frequency data. It is likely that a minimum of 10,000 ethnically matched control genomes are needed to obtain an accurate assessment of variant frequencies and thereby allow deductions to be made on phenotypic effects. Ongoing projects for example in the UK and China aim to sequence 100,000 and 1,000,000 control individuals respectively, to provide such accurate frequency estimates. Correct storage of these large datasets in a standardized format will be paramount for their utility in automated interpretation software. Initiatives like those from the Human Genome Variation Society [68] and the Global Alliance [69] try to raise awareness for these issues and engage large data centers to use open formats and facilitate open and efficient sharing of genetic and phenotypic data.

3.) STANDARDIZED AND QUANTITATIVE PHENOTYPIC DATA

In terms of availability and formats, phenotypic data has been relatively inaccessible to bioinformatics research. This is mostly because medical and/or phenotypic data is typically very unlike bioinformatics data (**Figure 3**). Medical data is often stored in non-digital formats, in the local language and usually based largely on subjective assessments rather than measurement based on objective classifications.

The value of standardized phenotypic information was demonstrated by the systematic phenotyping of knock-down/knock-out mice and the use of ontologies to systematically document phenotypic effects. This high level of standardization of these experiments has led to the successful implementation of mouse gene-phenotype information in bioinformatic models and variant effect prediction algorithms [31,70,71]. Efforts to standardize phenotypic data from human individuals like the human phenotype ontology (HPO) [72] are essential to enable bioinformatics to study phenotypic similarities in humans in a standardized way and to incorporate human phenotypic data in prediction models [73].

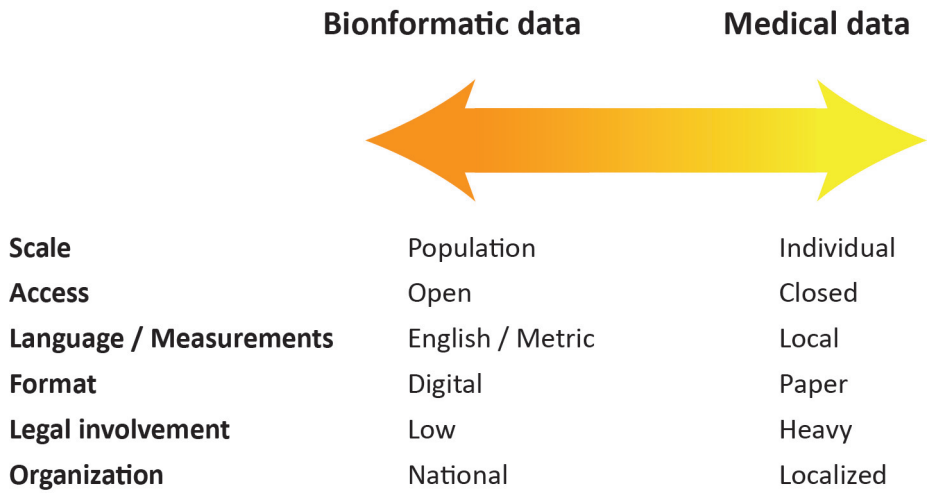


***Figure 3**. Typical differences between bioinformatic and medical data. Schematic overview of the characteristics of bioinformatic (left) and medical (right) data*

To date, most, if not all, existing databases storing phenotypic information lack the functionality to store quantitative measures. If a patient is for example described as macrocephalic, it is highly informative for computational models to know how many standard deviations this constitutes. The value of standardized quantitative phenotypic data in morphology related research has been previously shown [74]. That is, scientists performed 3D imaging on the faces of control individuals and individuals affected by different clinically recognized syndromes. Detailed quantitative data on facial characteristics such as nasal bridge width and eye spacing was used to determine and quantify differences between control and affected individuals [74,75]. This research illustrated the importance of recording and storing all phenotypes in a quantitative format: Studying the variability in the control groups allowed computer models to determine if patient phenotypes were significantly different from the control set and by what order of magnitude [74,75]. This method of assessing phenotypic information is less subjective than relying on the personal interpretation of a clinician.

Bioinformatics will have to play an important role in enabling clinicians to systematically store and share their phenotypic findings in a safe and easy way. There will need to be projects that develop intuitive user interfaces, in which clinicians can indicate which measurements have been performed, and compute (or indicate) the level of deviation from control individuals. Once readily available to each clinican/department, technologies like 3D imaging [74] will further reduce the

number of individual measurements a clinician needs to perform while providing a rich data set for bioinformatics algorithms and models.
To promote the use and reporting of quantitative phenotypic measurements it will be important that scientific publications are enforced to provide numeric measurements of the relevant phenotypes of both control and affected individuals. Once phenotypes are routinely stored in appropriate databases, it will be essential to link genetic variants with a relation to the phenotype to these databases.

4.) ACCURATE DESCRIPTIONS OF THE FUNCTIONAL REGIONS IN THE GENOME

Genome annotation involves the systematic description and linking of biological functions to regions or a genomic location. The first genome annotation was cytobanding, staining patterns enabled scientists to identify and sort chromosome pairs. At present the human genome is annotated for over 21.000 genes and increasingly complex features such as measures of evolution, transcription factor binding sites, and methylation patterns [76]. Genome annotations will continue to become richer, current functions will be defined more accurately and the function of more regions, especially non-genic regions, will be elucidated. For example, the ENCODE project has already made substantial progress in annotating the non-coding functional regions of the human genome [76]. These developments will result in large data sets, and as a consequence annotations will become a data source to reckon with. Bioinformatics will need to construct more efficient methods to store, version and distribute these if they are to be incorporated effectively into prediction models.
An important challenge in the interpretation of genetic variants will be to attribute appropriate weights to the different annotations/biological functions. That is, will variation in a region annotated to act as both a transcription factor binding site and a methylation site be more important for transcriptional activation or deactivation? Studies that evaluate the effect of variants in a systems-biology wide manner (DNA, RNA, protein, etc.) will provide biological read-outs used by bioinformaticians to construct appropriate weight matrices for different biological functions.
Annotations for evolutionary conservation [77,78] provide a numerical value representing how essential a particular piece of DNA or protein has been throughout evolution. These conservation values have been successfully incorporated into prediction algorithms to provide an estimate of variation tolerance [79,80]. Currently the most reliable and readily available conservation values are based on nucleotide conservation across species [77,78]. While this has proven an indicative measure for variant pathogenicity [**chapter 2**], it cannot always be calculated and has limited value for truncating variants and variants affecting naturally variable sites, such as “wobble bases”. Importantly, genetic plasticity has already been described to have

more far reaching effects than the wobble base effect, for example through RNA editing [81,82]. As we learn more about biological complexity, more of such caveats will be identified which will need to be incorporated in bioinformatics models. In addition, evolutionary conservation provides little value when studying the effect of variants in human or primate specific genes. Human-specific conservation values will become more accurate through large-scale sequencing of human control genomes, as these provide insights into human specific variability. Bioinformatics models can incorporate the amount of genetic variability seen in control genomes as a metric for variant tolerance of a certain gene or region. A recent publication for example demonstrated that ID associated genes are on average less tolerant for genetic variation in the general population compared to other genes [83].

5.) STANDARDIZED AND QUANTITATIVE FUNCTIONAL MEASUREMENTS

Bioinformatics has gained most from functional studies when these were performed systematically for a large set of genes in an organism like the knock-out studies performed in mice [31]. Ideally, variation at every position in the genome is evaluated for all types of functional consequences, and against all possible genetic backgrounds. This is, however, unattainable from both an economical and practical perspective, as the combinations are nearly endless. A more achievable aim is when researchers systematically evaluate (all) genetic variants in a predetermined set of functionally relevant regions and compare these to control specimens, as demonstrated recently for ID gene knock-out fly models [84].

Something which is often lacking in these knock-out studies is rescue experiments with human analogs which provides valuable information about relevance to the human phenotype [85]. While functional studies currently have little bioinformatic involvement at the experimental level, the correct interpretation and storage of the measurements already depends greatly upon bioinformatics [71,86,87]. Quantitative measures on variant effects require statistics to describe a variants effect in terms of strength and directionality. The involvement of bioinformatics in functional studies will increase as imaging techniques are used more often, as these require pattern recognition to identify and quantitate phenotypic effects. This was recently shown for a single gene *GATAD2B* [86] and more generally for a systematic screen of ID gene knock-outs [84].

Previous examples relate to knock-out models, however, mutations can also have antagonistic or synergistic effects which can also be studied in model organisms. A clinical example for which systematic functional studies of different variant types were successfully performed is Bardet-Biedl Syndrome [85]. For this disorder, researchers could use a relatively high throughput organism, the zebra fish, while still obtaining clinically relevant functional read-outs [85]. Using various functional

studies researchers were able to show that multiple variants within the same complex were required before a clinical phenotype became apparent. An additional observation was that single variants could result in sub-clinical phenotypes [85]. This study illustrates the importance of not only testing manually selected candidate variants for a phenotype but systematically testing all variants within a complex or pathway, as well as their combinations. Variant effect databases will need to correctly store and classify such combinatorial tests and their effects. Bioinformatics can then mine these databases and construct programs that learn from these functional studies and construct rules on which types of variants will likely influence each other.

6.) MORE DIVERSE AND ACCURATE VARIANT EFFECT PREDICTION

Bioinformatics played an instrumental role in the interpretation of variants detected by aCGH and MPS through a combination of prediction programs and annotations [25,31,33]. However, these successes were in a large part due to their applicability to a certain set of genes or a specific type of disorder, as shown in this thesis for sporadic ID [**chapter 2 & 3**].

Current protocols to predict the general effect size of any given non-synonymous single nucleotide variant still need to be improved immensely [88]. Improvements have been made by combining the outcomes of different prediction software tools, each relying on different (genomic/evolutionary) features [89]. There are, however, some fundamental flaws in these programs that reduce their usefulness and accuracy. The largest problems are that these programs are, i) unable to correctly analyze gain of function or dominant negative variants, ii) unable to predict the effect of non-coding variants, and iii) unable to predict the effect of different types of variation (SVs, synonymous SNVs, in-frame InDels, etc.). Bioinformatics will resolving several of these issues by constructing more sophisticated prediction models and combining data sources, like interaction networks and domain annotations with stability metrics as described below.

An important metric in current prediction tools is the effect a variant has on stability or functionality of the affected element. For variants in the coding region, this is, mostly predicted based on the putative effect on protein stability. Accurate measurements on the change in protein stability and function require accurate 3D models of the natural state and function of the affected protein. For example the effect prediction of variants involved in Cantú syndrome, affecting *ABCC9* [90,91], can only be done correctly when taking into account the presence and flux of potassium through the ion channel. The correct modeling of the biochemical function of proteins will allow prediction tools to more accurately predict the effect on stability of a variant.

Predicting variant effects outside the coding region is an enormous challenge at present. Factors such as local DNA structure or sequence motifs [92] will be relevant metrics to estimate variant effects. A sequence motif captures the characteristics of a certain functional unit through sequence comparisons [93], usually in an evolutionary or genome wide perspective. Sequence motifs can be used to compute the degree of disruption of the element and secondly to define similar regions in the human genome to measure their variation tolerance in control genomes. Such region or pattern based methods will also help to interpret other forms of genetic variation. Rather than looking at a single nucleotide, predictions can be based on entire genomic regions. To predict the effect of complex events like SVs a correct reconstruction of the patient's genome is essential to perform such region based analysis. For example, prediction programs need information on whether duplication events are situated in tandem or distal to each other, in order to characterize the disrupted regions and newly introduced elements. Local *de novo* assembly is a particularly useful bioinformatics tool in resolving these issues and will be more powerful as longer sequence reads become available.

7.) ASSESSMENT OF VARIANT EFFECT ACROSS DIFFERENT LEVELS OF BIOLOGY

An individual's phenotype is not just the sum of the effects of each variant individually; there is an intricate interplay between genes, proteins and their regulatory mechanisms. Accounting for the biological complexity of genetic regulation and interaction will be a major challenge in constructing accurate variant effect prediction models.

Figure 4 depicts an example of a network containing both proteins and metabolites. In this network there are several proteins which are known to play a role in ID when mutated (in red). Effect prediction models will need data on the flow of signals and metabolites through the pathway and a complete overview of the interactions and their directions. Once these data are available, modeling can be performed in such a network to estimate if perturbations of different proteins have similar effects on the network. For example, one would like to investigate if another protein can counteract the perturbation through regulatory loops. When networks are detailed enough to answer these questions personalized medicine will start to become possible, as has been shown by preliminary successes for certain drugs [94]. For instance, for Warfarin, an anticoagulant used to prevent thrombosis, relative simple statistics of genotype and dosage effect correlation can now be used to provide a personalized medication advise for this drug [94].

To predict drug and variant effects throughout the body and across the different levels of biology will require high quality interaction, expression and metabolic networks. The construction of interaction and correlation networks of genes/

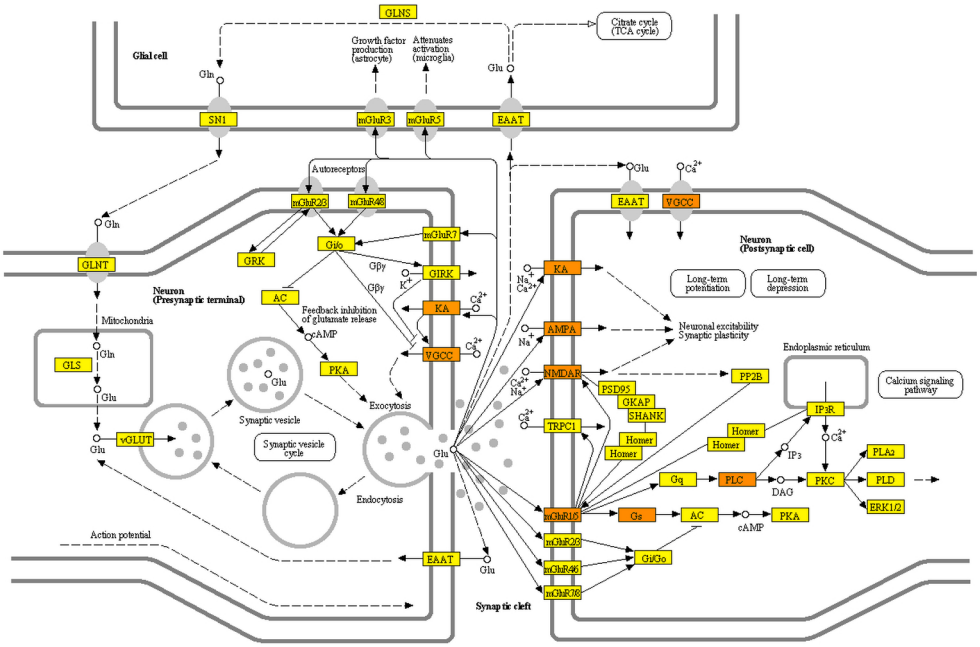


***Figure 4**, Glutamergic synapse pathway from KEGG [98]. Components highlighted in orange are genes/proteins known to be mutated in individuals with ID [100,101].*

proteins across different cell types, and stages of development, will require both detailed studies, in a single organ, as well as more general genomic and proteomic approaches [95]. Genome wide experiments in different tissues from healthy adult individuals will provide the backbone for these networks in terms of connectivity and directionality [95,96].

When phenotypic predictions need to me made about germline variants, the developmental processes in which the affected gene/genomic region is involved will form an essential part of the effect prediction model. Developmental studies will require more sophisticated experiments to capture the changes through time, which will limit the amount of available data [97]. Nonetheless, developmental experiments will provide the insights needed to construct models on how networks and interactions adapt throughout development and provide the data to capture such biological complexity in computer models. Fundamental principles in bioinformatics such as network and graph theory [2,98,99] will be instrumental in the construction and use of these complex networks in variant effect prediction.

8.) COMBINED EFFECT PREDICTIONS ACROSS MULTIPLE VARIANTS

When sufficient data is available about the quantitative effect of a variant and the disrupted mechanisms, the final, and likely the toughest, challenge for bioinformatics will be the combined interpretation of all variants. In this process, all predictions of individual variants will need to be combined and weighted correctly to predict the net effect of all variation on the phenotype. Any assumptions or approximations made in previous steps will propagate into this combinational step and can be influenced in such a way that the outcome differs by orders of magnitude from the actual effect. It will be key to;

1. quantify the disruptive effect on the system
2. understand and quantify the directionality of the flow within the system.

In **Figure 2**, part of the classification relates to risks and modifiers, variants that have a milder or environment dependent effect. It is hard to predict the effect of such a variant by itself; it requires knowledge on genetic background and possibly environmental factors. Quantitative read-outs of variant effects will help interpret these modifiers and to consider their load in the complete model.

The importance of correctly assessing the net effect and combining the variants in prediction models was demonstrated by case studies in Bardet-Biedl Syndrome [85]. Despite having an appropriate functional read-out of the variant effects, the sum of the effects of every single variant was not equal to the effect in the specimens harboring the variants [85]. These combinatorial analyses will not only require a lot of computational power but also many connections between different data sources [102].

OTHER CHALLENGES RELATED TO BIOINFORMATICS AND GENETICS

Apart from the obvious role of bioinformatics in improving variation interpretation there are several other, more general, issues related to genomic experiments which I have not yet discussed in this thesis while they are important topics in current (political) debates [69,102]. Firstly, an ever changing and very important topic in human related research is privacy [103]. It has been shown that privacy cannot be guaranteed for rich data (in terms of identifiable features) such as genomic profiles [104,105]. It is important to realize this caveat and adjust consent accordingly as has been suggested by the “Personal Genome Project” and others [66,102,103]. Informatics will play an important role in data security, and thereby privacy, through encryption and access rights. It should, however, be noted that for the years to come (bioinformatics) research will benefit most readily from data available through open consent coupled to proper anti-discrimination legislation [102].

A second opportunity for bioinformatics, is the need for interfaces that explain genetic testing results to counselors and patients. Bioinformatics will continue to

play an essential role in bringing concepts such as risk and chance on disease to the field of personalized medicine. Companies like 23andMe present examples of how disease risks can be made understandable to the general population through figures and statistics. I envision that there will be an increase in (bioinformatic) companies developing and offering interpretation of genetic variants in terms of health and disease, but also in related fields such as match making and pre-conception genetics.

CLOSING REMARKS AND A BRIEF OUTLOOK

There are several points which have been discussed which I think will be crucial to the success of bioinformatics in future genetic research. The first is a point raised by many others [2,6,69,102], the need for standards and databases for both variants and phenotypes. Without these bioinformatics will be severely limited in its potential to link data sources and make correct inferences about genotype-phenotype correlations. The second important issue is the need to realize that there are very few variants that will have the exact same phenotypic outcome in different individuals. Factors like ethnicity, epigenetics, regulation, expression levels and interactions are all part of the final equation and these types of data need to be gathered, stored and shared soon. This brings me to the next point, the need to share; many labs have excellent data about a certain pathway or a set of genes, if this data is made freely available to others I am confident that the field will progress even quicker.

This discussion started with the assumption that sequencing and variant identification will become trivial. In light of recent developments and the promise of emerging techniques such as single cell and nanopore sequencing [37,59,106], I think this assumption will become reality within the next five years. The next step is interpretation, to achieve a completely systematic and accurate interpretation of genetic variation, prediction models will ideally encompass a computer model of the complete human body [26,107]. While this may not be reality within five years from now, we may be able to unravel the underlying biological pathways of many diseases and start targeted treatment trials [108,109]. For certain genes and pathways we will be able to accurately identify and compute the effects of variation and understand their roles in disease [110]. Moreover, for heterogeneous disorders, such as ID, we will have a better understanding of the underlying genetic defects and their corresponding (sub) phenotypes allowing for more fine-grained studies on the underlying biochemical mechanisms and potentially studies aiming at gene or pathway specific treatment options.

A question which remains is whether advances in informatics, genetics, data sharing and bioinformatics, combined with new technologies such as nanopore and single cell sequencing will enable a bedside genome analyzer [37,106]. Such a machine

will sequence and interpret an individual's genome in real time for a low price; while I am confident that this will become a reality, I do not expect it to happen within the next five years. Nonetheless, bioinformatics will continue to play an important role in bringing all the technological advances in the 'omics' fields together and combine these into personalized medicine, by linking data sources and providing comprehensive interpretation of variant effects, and finally in presenting these to researchers, clinicians and patients in an understandable and meaningful way.

REFERENCES

1. Mendel, G. Gregor Mendel's letters to Carl Nägeli, 1866-1873. Genetics 35, 1-29 (1950).
2. Searls, D. B. The roots of bioinformatics. PLoS Comput. Biol. 6, e1000809 (2010).
3. Crow, J. F. A diamond anniversary: the first chromosomal map. Genetics 118, 1-3 (1988).
4. Haldane, J. B. S. The rate of spontaneous mutation of a human gene. J. Genet. 31, 317-326 (1935).
5. Rider, R. E. A mathematician: alan turing. Science 223, 807 (1984).
6. Ouzounis, C. A. & Valencia, A. Early bioinformatics: the birth of a discipline--a personal view. Bioinformatics 19, 2176-90 (2003).
7. Shrödinger, E. (1946) What is life? the physical aspect of the living cell. New York: MacMillan. 91 p.
8. Fisher, R. A. Gene frequencies in a cline determined by selection and diffusion. Biometrics 6, 353-61 (1950).
9. Fisher R. A. (1930) The genetical theory of natural selection. Variorum edition, 2000. New York: Oxford University Press. 318 p.
10. Watson, J. D. & Crick, F. H. C. Genetic implications of the structure of deoxyribonucleic acid. Nature 171, 964-967 (1953).
11. Nature Archives. double helix 50 years of DNA. (2003). at <http://www.nature.com/nature/dna50/index.html>
12. Ouzounis, C. A. Two or three myths about bioinformatics. Bioinformatics 16, 187-9 (2000).
13. Gerstein, M. B. et al. What is a gene, post-ENCODE? History and updated definition. Genome Res. 17, 669-81 (2007).
14. Anderson HL (1986) Metropolis, Monte Carlo, and the MANIAC. Los Alamos Science 14: 96-108. Available: http://library.lanl.gov/cgi-bin/getfile?00326886.pdf.
15. Adams J (2008) Sequencing human genome: the contributions of Francis Collins and Craig Venter. Nature Education 1(1). Available: http://www.nature.com/scitable/topicpage/Sequencing-Human-Genome-the-Contributions-of-Francis-686.
16. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443-53 (1970).
17. Gibbs, A. J. & McIntyre, G. A. The diagram, a method for comparing sequences. Its use

with amino acid and nucleotide sequences. Eur J Biochem. 16, 1-11 (1970).

18. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463-7 (1977).
19. Ouzounis, C. A. Rise and demise of bioinformatics? Promise and progress. PLoS Comput. Biol. 8, e1002487 (2012).
20. Van Rossum, T., Tripp, B. & Daley, D. SLIMS--a user-friendly sample operations and inventory management system for genotyping labs. Bioinformatics 26, 1808-10 (2010).
21. Ewald, R., Himmelspach, J., Jeschke, M., Leye, S. & Uhrmacher, A. M. Flexible experimentation in the modeling and simulation framework JAMES II--implications for computational systems biology. Brief. Bioinform. 11, 290-300 (2010).
22. Gauthier, J. et al. *De novo* mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. Proc. Natl. Acad. Sci. U. S. A. 107, 7863-8 (2010).
23. Hehir-Kwa, J. Y. et al. Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. DNA Res. 14, 1-11 (2007).
24. Pinto, D. et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. 29, 512-20 (2011).
25. Horner, D. S. et al. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief. Bioinform. 11, 181-97 (2010).
26. Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J. & Altman, R. B. Bioinformatics challenges for personalized medicine. Bioinformatics 27, 1741-8 (2011).
27. Moore, G. E. Cramming More Components onto Integrated Circuits. Proc. IEEE 86, 82-85 (1998).
28. Rothberg, J. M. et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348-52 (2011).
29. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. Science 337, 1628 (2012).
30. De Leeuw, N. et al. SNP array analysis in constitutional and cancer genome diagnostics -copy number variants, genotyping and quality control. Cytogenet. Genome Res. 135, 212-21 (2011).
31. Hehir-Kwa, J. Y. et al. Accurate distinction of pathogenic from benign CNVs in mental retardation. PLoS Comput. Biol. 6, e1000752 (2010).
32. Kanehisa, M. & Bork, P. Bioinformatics in the post-sequence era. Nat. Genet. 33 Suppl, 305-10 (2003).
33. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. Eur. J. Hum. Genet. 20, 490-7 (2012).
34. Hehir-Kwa, J., Pfundt, R., Veltman, J. & de Leeuw, N. Pathogenic or not? Assessing the clinical relevance of copy number variants. Clin. Genet. (2013). doi:10.1111/cge.12242
35. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing.

Bioinformatics 28, 2711-8 (2012).

36. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-303 (2010).
37. Ku, C.-S. & Roukos, D. H. From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. Expert Rev. Med. Devices 10, 1-6 (2013).
38. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. Nat. Methods 8, 61-5 (2011).
39. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. Science 330, 641-6 (2010).
40. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363-76 (2011).
41. Guerrier, S. et al. The F-BAR domain of *SRGAP2* induces membrane protrusions required for neuronal migration and morphogenesis. Cell 138, 990-1004 (2009).
42. Dennis, M. Y. et al. Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. Cell 149, 912-22 (2012).
43. Charrier, C. et al. Inhibition of *SRGAP2* function by its human-specific paralogs induces neoteny during spine maturation. Cell 149, 923-35 (2012).
44. Mefford, H. C. & Eichler, E. E. Duplication hotspots, rare genomic disorders, and common disease. Curr. Opin. Genet. Dev. 19, 196-204 (2009).
45. Sharp, A. J. et al. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat. Genet. 38, 1038-42 (2006).
46. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7, e47768 (2012).
47. Ewen, K. R. et al. Identification and Analysis of Error Types in High-Throughput Genotyping. Am. J. Hum. Genet. 67, 727-736 (2000).
48. Pang, A. W. et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 11, R52 (2010).
49. Ramu, A. et al. DeNovoGear: *de novo* indel and point mutation discovery and phasing. Nat. Methods (2013). doi:10.1038/nmeth.2611
50. Stewart, C. et al. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. 7, e1002236 (2011).
51. Okoniewski, M. J. et al. Precise breakpoint localization of large genomic deletions using PacBio and Illumina next-generation sequencers. Biotechniques 54, 98-100 (2013).
52. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1, 18 (2012).
53. Chu, T.-C. et al. Assembler for *de novo* assembly of large genomes. Proc. Natl. Acad. Sci. U. S. A. (2013). doi:10.1073/pnas.1314090110
54. Bailey, J. A. et al. Recent segmental duplications in the human genome. Science 297, 1003-7 (2002).
55. Limaye, N., Boon, L. M. & Vikkula, M. From germline towards somatic mutations in the

pathophysiology of vascular anomalies. Hum. Mol. Genet. 18, R65-74 (2009).

56. Campbell, P. J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. 40, 722-9 (2008).
57. Erickson, R. P. Somatic gene mutation and human disease other than cancer: an update. Mutat. Res. 705, 96-106 (2010).
58. Tartaglia, M. et al. Genetic evidence for lineage-related and differentiation stage-related contribution of somatic PTPN11 mutations to leukemogenesis in childhood acute leukemia. Blood 104, 307-13 (2004).
59. Wang, J., Fan, H. C., Behr, B. & Quake, S. R. Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. Cell 150, 402-12 (2012).
60. McConnell, M. J. et al. Mosaic copy number variation in human neurons. Science 342, 632-7 (2013).
61. Ashley, E. A. et al. Clinical assessment incorporating a personal genome. Lancet 375, 1525-35 (2010).
62. Kojima, K. et al. A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads. Bioinformatics (2013). doi:10.1093/bioinformatics/btt503
63. Abecasis, G. R. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061-73 (2010).
64. Boomsma, D. I. et al. The Genome of the Netherlands: design, and project goals. Eur. J. Hum. Genet. (2013). doi:10.1038/ejhg.2013.118
65. Lupski, J. R., Belmont, J. W., Boerwinkle, E. & Gibbs, R. a. Clan genomics and the complex architecture of human disease. Cell 147, 32-43 (2011).
66. Church, G. M. The personal genome project. Mol. Syst. Biol. 1, 2005.0030 (2005).
67. Arias, A. et al. Molecular dissection of a viral quasispecies under mutagenic treatment: positive correlation between fitness loss and mutational load. J. Gen. Virol. 94, 817-30 (2013).
68. HGVS. at <http://www.hgvs.org/>
69. Global Alliance. Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data. (2013). at <http://oicr.on.ca/files/public/White_paper_2013_06_03_FINAL.pdf>
70. Chen, J., Xu, H., Aronow, B. J. & Jegga, A. G. Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics 8, 392 (2007).
71. Webber, C. et al. Forging links between human mental retardation-associated CNVs and mouse gene knockout models. PLoS Genet. 5, e1000531 (2009).
72. Robinson, P. N. & Mundlos, S. The human phenotype ontology. Clin. Genet. 77, 525-34 (2010).
73. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 37, W305-11 (2009).

74. Hammond, P. & Suttie, M. Large-scale objective phenotyping of 3D facial morphology. Hum. Mutat. 33, 817-25 (2012).

75. Hammond, P. et al. Fine-grained facial phenotype-genotype analysis in Wolf-Hirschhorn syndrome. Eur. J. Hum. Genet. 20, 33-40 (2012).

76. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636-40 (2004).

77. Cooper, G. M. et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15, 901-13 (2005).

78. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110-21 (2010).

79. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073-81 (2009).

80. Ng, P. C. {SIFT:} predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812-3814 (2003).

81. Chepelev, I., Wei, G., Tang, Q. & Zhao, K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res. 37, e106 (2009).

82. Shah, S. P. et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 461, 809-13 (2009).

83. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 9, e1003709 (2013).

84. Oortveld, M. A. W. et al. Human Intellectual Disability Genes Form Conserved Functional Modules in Drosophila. PLoS Genet. 9, e1003911 (2013).

85. Ashkinadze, E., Rosen, T., Brooks, S. S., Katsanis, N. & Davis, E. E. Combining fetal sonography with genetic and allele pathogenicity studies to secure a neonatal diagnosis of Bardet-Biedl syndrome. Clin. Genet. 83, 553-9 (2013).

86. Willemsen, M. H. et al. *GATAD2B* loss-of-function mutations cause a recognisable syndrome with intellectual disability and are associated with learning deficits and synaptic undergrowth in Drosophila. J. Med. Genet. 50, 507-14 (2013).

87. Zaghloul, N. A. et al. Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. Proc. Natl. Acad. Sci. U. S. A. 107, 10602-7 (2010).

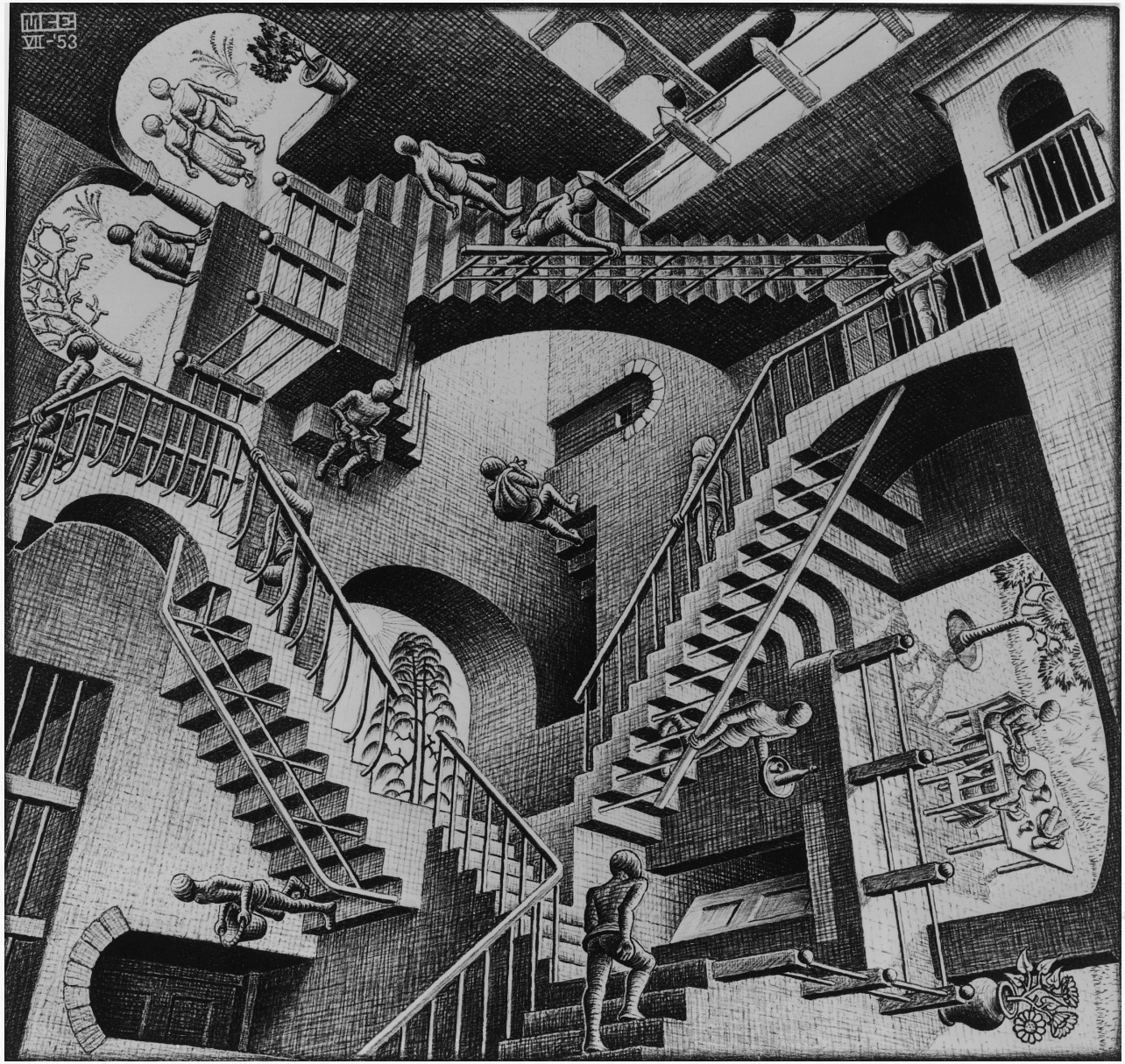
88. Gray, V. E., Kukurba, K. R. & Kumar, S. Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. Bioinformatics 28, 2093-6 (2012).

89. González-Pérez, A. & López-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. 88, 440-9 (2011).

90. Harakalova, M. et al. Dominant missense mutations in *ABCC9* cause Cantú syndrome. Nat. Genet. 44, 793-6 (2012).

91. van Bon, B. W. M. et al. Cantú syndrome is caused by mutations in *ABCC9*. Am. J. Hum. Genet. 90, 1094-101 (2012).

92. Bateman, A. et al. The Pfam protein families database. Nucleic Acids Res. 30, 276-80 (2002).
93. Eddy, S. R. Profile hidden Markov models. Bioinformatics 14, 755-63 (1998).
94. Klein, T. E. et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. N. Engl. J. Med. 360, 753-64 (2009).
95. Hsiao, L. L. et al. A compendium of gene expression in normal human tissues. Physiol. Genomics 7, 97-104 (2001).
96. Pardo, L. M. et al. Regional differences in gene expression and promoter usage in aged human brains. Neurobiol. Aging 34, 1825-36 (2013).
97. Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. Nature 478, 483-9 (2011).
98. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27-30 (2000).
99. Ronen, M., Rosenberg, R., Shraiman, B. I. & Alon, U. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. Proc. Natl. Acad. Sci. U. S. A. 99, 10555-60 (2002).
100. van Bokhoven, H. Genetic and epigenetic networks in intellectual disabilities. Annu. Rev. Genet. 45, 81-104 (2011).
101. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, M. Online Mendelian Inheritance in Man, OMIM®. (2013). at <http://omim.org/>
102. Schadt, E. E. The changing privacy landscape in the era of big data. Mol. Syst. Biol. 8, 612 (2012).
103. Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. From genetic privacy to open consent. Nat. Rev. Genet. 9, 406-11 (2008).
104. Schadt, E. E., Woo, S. & Hao, K. Bayesian method to predict individual SNP genotypes from gene expression data. Nat. Genet. 44, 603-8 (2012).
105. Homer, N. et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 4, e1000167 (2008).
106. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. Hum. Mol. Genet. 19, R227-40 (2010).
107. Lindblom, A. & Robinson, P. N. Bioinformatics for human genetics: promises and challenges. Hum. Mutat. 32, 495-500 (2011).
108. Limdi, N. A. & Veenstra, D. L. Warfarin pharmacogenetics. Pharmacotherapy 28, 1084-97 (2008).
109. Thorn, C. F., Klein, T. E. & Altman, R. B. Pharmacogenomics and bioinformatics: {PharmGKB}. Pharmacogenomics 11, 501-505 (2010).
110. Gage, B. F. & Lesko, L. J. Pharmacogenetics of warfarin: regulatory, scientific, and clinical issues. J. Thromb. Thrombolysis 25, 45-51 (2007).


MCE
VII-53

Chapter 7

Summary / Samenvatting

SUMMARY

In this thesis the detection and role of newly occurring (*de novo*) mutations are studied in relation to ID. ID is typically defined by an intelligence quotient (IQ) below 70 and (severe) developmental delays in adaptive behavior, starting before the age of 18 years. The research was focused on sporadic cases, with no familial history for the disorder. It has long been thought that the sporadic nature of ID might be explained by the occurrence of *de novo* germline mutations, referring to errors that occur in the genetic material during the formation of the egg- or sperm cell. There was however no supporting evidence for this hypothesis due to technical limitations for the detection of these mutations.

The availability of new genetic techniques allowed researchers to study *de novo* mutations genome wide (**Chapter 1**). It has been shown that *de novo* mutations can play an important role in both rare and common disorders. The extent to which *de novo* mutations impact the occurrence of human disease has been shown to depend on both genetic factors (like target mutability and size) and other factors such as parental age. The interpretation of the phenotypic effect of *de novo* mutations remains challenging and requires combinations of deep phenotyping, statistical analysis, functional studies and recurrence testing in additional cohorts of patients with ID.

The work described in this thesis focused mainly on sequencing the protein coding regions (the exome) of the human genome. Exome sequencing of 10 ID patients and their unaffected parents revealed that small (one to several base pairs) *de novo* mutations were the probable cause of ID in 6 patients (**Chapter 2**). The mutations in these patients occurred in genes involved in brain development and were predicted to impact protein function. These data showed that exome sequencing combined with a family-based experimental setup is a reliable method for the detection of *de novo* mutations.

As our initial study suggested implications for preventive and diagnostic strategies in ID, we further evaluated the diagnostic utility of exome sequencing in 100 patients with severe unexplained ID, and their parents (**Chapter 3**). These patients had all reached the end stage of routine diagnostic procedures. A likely genetic cause of ID was identified in 16% of the patients, mostly based on *de novo* mutations. Additionally, in 19 patients *de novo* mutations were identified in candidate ID genes. We therefore postulated that the total diagnostic yield of this technique could be as high as 35% in this group of patients, and therefore warrants its implementation in routine diagnostics.

Apart from evaluating the ability of exome sequencing to identify small coding mutations we also investigated its ability to detect larger mutations in the form of CNVs (**Chapter 4**). CNVs delete or duplicate regions in the genome, thereby altering

the number of copies of that particular piece of DNA. The occurrence of *de novo* CNVs currently explains 10-20% of ID cases and these mutations are routinely detected using genomic microarrays. Their detection from sequencing data would be a valuable asset. A selection of 10 patients with 12 known, ID-related CNVs was used to assess whether these CNVs could be detected from exome sequencing data. To this end, four prediction programs were evaluated of which one was able to identify 11 of the 12 CNVs. From these data, we predict that 88% of all diagnostically relevant *de novo* CNVs (containing at least 3 exons) would be detectable from exome sequencing data, further increasing the utility of this technique in ID diagnostics.

Finally the novel sequencing technique was used in a genome wide manner to detect *de novo* chromosomal aneuploidies (**Chapter 5**) in a prenatal setting. Prenatal testing of fetal DNA for aneuploidies occurs mostly through invasive procedures, which have an intrinsic risk of miscarriage. Using a non-invasive test would circumvent this risk. Blood of pregnant women was used to determine whether chromosomal aneuploidies of the fetus could be reliably detected using whole genome massive parallel sequencing. Using this technique we were able to reliably detect copy number changes of whole chromosomes, suggesting that this test can readily be implemented in a prenatal setting, especially for chromosomes 13, 18, 21.

Based on the increased diagnostic yield and the potential to reduce of the number of genetic tests performed per patient exome sequencing has now been included in the routine diagnostic strategy of patients coming to our clinic. The implementation of non-invasive prenatal trisomy testing is also well underway in The Netherlands because of the advantages mentioned above. Collectively, these examples illustrate the added diagnostic value of next generation sequencing approaches s in comparison to older techniques. With costs of sequencing decreasing in combination with its other advantages, it may be anticipated that next generation sequencing approaches will become the first tier genetic test for many diagnostic applications. However, the clinical interpretation of mutations identified by such tests is complex and requires a lot of multidisciplinary research.

Our current knowledge of the human genome remains limited as we only now start to understand the function of a small percentage of the genome, mostly how disruptions in protein-coding parts can lead to disease. To further improve predictions of (*de novo*) mutations (**Chapter 6**) requires a better understanding of the genome and its variation in healthy individuals. Mutation effect prediction models will benefit from the inclusion of (tissue-specific) protein- and expression data. Also, linking genetic variation to (clinical) phenotypes and other biological data will be greatly substantiated by recording phenotypes in a standardized and quantitative manner. This together will eventually lead to a prediction of the combined effect of all genetic variation in a single model rather that each mutation separately.

SAMENVATTING

In dit proefschrift is de detectie en de rol van nieuw ontstane (*de novo*) mutaties onderzocht in relatie tot het ontstaan van een verstandelijke beperking (VB). VB wordt getypeerd door een intelligentie quotiënt (IQ) onder de 70 en (ernstige) ontwikkelingsstoornissen in adaptief gedrag, voor het 18de levensjaar. Het onderzoek was gericht op sporadische gevallen, zonder familiaire voorgeschiedenis voor de aandoening. Er is lang gedacht dat het sporadisch voorkomen van VB verklaart kon worden door *de novo* mutaties; fouten die ontstaan in het genetische materiaal tijdens de vorming van de ei- of zaadcel. Er was echter geen bewijs voor deze hypothese door technische beperkingen voor de detectie van deze mutaties.

De beschikbaarheid van nieuwe genetische technieken stelde onderzoekers in staat om *de novo* mutaties genoomwijd te bestuderen (**hst. 1**). Er werd aangetoond dat *de novo* mutaties een belangrijke rol kunnen spelen in het ontstaan van zowel zeldzame als vaak voorkomende aandoeningen. De mate waarin *de novo* mutaties invloed hebben op het ontstaan van ziekte is afhankelijk van zowel genetische factoren (de grootte en stabiliteit van het doelwit) als andere factoren zoals de leeftijd van de ouders. De interpretatie van het fenotypische effect van *de novo* mutaties blijft uitdagend en vereist combinaties van diepe fenotypering, statistische analyses, functionele testen en herhalingstesten in andere VB patiënten.

Het in dit proefschrift beschreven onderzoek richtte zich vooral op het sequencen van de eiwit coderende gebieden (het exoom) van het menselijk genoom. Exoom sequencing van 10 VB patiënten en hun gezonde ouders toonde aan dat kleine (een tot enkele basenparen) *de novo* mutaties de waarschijnlijke oorzaak waren van VB in 6 patiënten (**hst. 2**). De mutaties vonden plaats in genen die betrokken zijn bij de ontwikkeling van de hersenen en werden voorspeld de eiwit functie te beïnvloeden. De gegevens toonden aan dat exoom sequencing in zowel de patiënt als de ouders een betrouwbare methode is voor de detectie van *de novo* mutaties.

Omdat de eerste studie implicaties toonde voor preventieve en diagnostische strategieën in VB, evalueerden we de diagnostische bruikbaarheid bij 100 patiënten met ernstige onverklaarbare VB en hun ouders (**hst. 3**). Deze patiënten hadden het eindstadium van routine diagnostische procedures bereikt. Een waarschijnlijke genetische oorzaak van VB werd geïdentificeerd in 16% van de patiënten, meestal gebaseerd op *de novo* mutaties. Bovendien werden in 19 patiënten *de novo* mutaties geïdentificeerd in kandidaat VB genen. We verwachtten daarom dat de totale diagnostisch opbrengst van deze techniek zo hoog als 35% kan zijn in deze groep patiënten, en rechtvaardigt de toepassing in routine diagnostiek.

Naast het vermogen van exoom sequencing om kleine coderende mutaties te identificeren onderzochten we ook het vermogen om grote veranderingen te detecteren in de vorm van CNVs (**hst. 4**). CNVs verwijderen of dupliceren regio's in het genoom, waardoor het aantal exemplaren van dat stuk DNA veranderd.

Het optreden van *de novo* CNVs verklaart momenteel 10-20% van VB gevallen, en wordt gedetecteerd met behulp van genomische microarrays. CNV detectie uit sequencing data zou een waardevolle toevoeging zijn. Een selectie van 10 patiënten met 12 bekende, VB gerelateerde, CNVs werd gebruikt om te beoordelen of CNVs konden worden gedetecteerd met exoom sequencing. Hiertoe werden vier programma's getest waarvan eentje 11 van de 12 CNVs detecteerde. Uit deze gegevens voorspelden wij dat 88% van diagnostisch relevante *de novo* CNVs (die 3 of meer exonen bevatten) door exoom sequencing gedetecteerd kan worden, een verdere verbetering van deze techniek voor VB diagnostiek.

Tenslotte werd de nieuwe sequencing techniek prenataal gebruikt om op een genoom wijde manier *de novo* chromosomale aneuploïdieën te detecteren (**hst. 5**). Prenataal onderzoek van foetaal DNA voor aneuploïdieën gebeurt meestal met invasieve procedures, die een intrinsiek risico op een miskraam hebben. Het gebruik van een niet-invasieve test zou dit risico omzeilen. Bloed van zwangere vrouwen werd gebruikt om te bepalen of chromosomale aneuploïdieën van de foetus betrouwbaar konden worden gedetecteerd met behulp van genoom sequencing. Met deze techniek konden we betrouwbaar duplicaties van gehele chromosomen detecteren, wat suggereerde dat deze test geïmplementeerd kan worden in prenataal onderzoek, specifiek voor chromosomen 13, 18, 21.

Op basis van de verhoogde opbrengst en de mogelijkheid van het terugbrengen van het aantal genetische tests per patiënt is exoom sequencing opgenomen in de routine diagnostiek in onze kliniek. De implementatie van niet-invasieve prenatale testen is in volle gang in Nederland vanwege de eerder genoemde voordelen. Tezamen illustreren deze voorbeelden de toegevoegde waarde ten opzichte van oudere technieken. De afnemende kosten van sequencing in combinatie met de andere voordelen wekt de verwachting dat sequencing de standaard genetische test wordt voor alle genetische aandoeningen. De klinische interpretatie en rapportage van mutaties geïdentificeerd door een dergelijke test is complex en vereist multi-disciplinaair onderzoek.

Onze huidige kennis van het menselijk genoom is nog zeer beperkt, we beginnen een klein percentage van het genoom te begrijpen, met name hoe verstoringen in eiwit coderende delen kunnen leiden tot ziekte. Voorspellen of een (*de novo*) mutatie schadelijk is (**hst. 6**) vereist een beter begrip van het genoom en de variatie in gezonde individuen. Mutatie effect voorspellingsmodellen zullen profiteren van het gebruiken van (weefsel specifieke) eiwit en expressie data. Ook het koppelen van genetische variatie aan (klinische) fenotypes en andere biologische gegevens zal sterk worden ondersteund door het gestandaardiseerd en kwantitatief opslaan van fenotypes. Dit samen zal uiteindelijk leiden tot een voorspelling van het gecombineerde effect van alle genetische variatie in een enkel model in plaats van elke mutatie afzonderlijk.

List of publications

Bosch DG, Boonstra FN, Gonzaga-Jauregui C, Xu M, **de Ligt J**, Jhangiani S, Wiszniewski W, Muzny DM, Yntema HG, Pfundt R, Vissers LE, Spruijt L, Blokland EA, Chen CA; Baylor-Hopkins Center for Mendelian Genomics, Lewis RA, Tsai SY, Gibbs RA, Tsai MJ, Lupski JR, Zoghbi HY, Cremers FP, de Vries BB, Schaaf CP. *NR2F1* Mutations Cause Optic Atrophy with Intellectual Disability. Am J Hum Genet. 2014; doi: 10.1016/j.ajhg.2014.01.002. [Epub ahead of print]

Buysse K, **de Ligt J,** Janssen IM, van Bon BW, Gomes I, Hehir-Kwa JY, Eggink AJ, van Vugt JM, Vissers LE, Geurts van Kessel A, Faas BH. Detecting fetal subchromosomal aberrations by MPS: an unexpected discrepancy between amniocyte DNA and ccffDNA. Prenat Diagn. 2014; doi: 10.1002/pd.4312. [Epub ahead of print]

Buysse K, Beulen L, Gomes I, Gilissen C, Keesmaat C, Janssen IM, Derks-Willemen JJ, **de Ligt J**, Feenstra I, Bekker MN, van Vugt JM, Geurts van Kessel A, Vissers LE, Faas BH. Reliable noninvasive prenatal testing by massively parallel sequencing of circulating cell-free DNA from maternal plasma processed up to 24h after venipuncture. Clin Biochem. 2013; 46(18):1783-6.

Schuurs-Hoeijmakers JH, Vulto-van Silfhout AT, Vissers LE, van de Vondervoort II, van Bon BW, **de Ligt J**, Gilissen C, Hehir-Kwa JY, Neveling K, del Rosario M, Hira G, Reitano S, Vitello A, Failla P, Greco D, Fichera M, Galesi O, Kleefstra T, Greally MT, Ockeloen CW, Willemsen MH, Bongers EM, Janssen IM, Pfundt R, Veltman JA, Romano C, Willemsen MA, van Bokhoven H, Brunner HG, de Vries BB, de Brouwer AP. Identification of pathogenic gene variants in small families with intellectually disabled siblings by exome sequencing. J Med Genet. 2013; 50(12):802-11.

de Ligt J*, Boone PM*, Pfundt R, Vissers LE, Richmond T, Geoghegan J, O'Moore K, de Leeuw N, Shaw C, Brunner HG, Lupski JR, Veltman JA, Hehir-Kwa JY. Detection of clinically relevant copy number variants with whole-exome sequencing. Hum Mutat. 2013; 34(10):1439-48.

Willemsen MH, Nijhof B, Fenckova M, Nillesen WM, Bongers EM, Castells-Nobau A, Asztalos L, Viragh E, van Bon BW, Tezel E, Veltman JA, Brunner HG, de Vries BB, **de Ligt J**, Yntema HG, van Bokhoven H, Isidor B, Le Caignec C, Lorino E, Asztalos Z, Koolen DA, Vissers LE, Schenck A, Kleefstra T. *GATAD2B* loss-of-function mutations cause a recognisable syndrome with intellectual disability and are associated with learning deficits and synaptic undergrowth in Drosophila. J Med Genet. 2013; 50(8):507-14.

de Ligt J, Veltman JA, Vissers LE. Point mutations as a source of *de novo* genetic disease. Curr Opin Genet Dev. 2013; 23(3):257-63.

Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, Boomsma DI, van Duijn CM, Slagboom PE, van Ommen GJ, Wijmenga C; **Genome of the Netherlands Consortium,** de Bakker PI, Sunyaev SR. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet. 2013; 9(2):e1003301.

de Ligt J*, Willemsen MH*, van Bon BW*, Kleefstra T*, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, Scheffer H, de Vries BB, Brunner HG, Veltman JA, Vissers LE. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012; 367(20):1921-9.

Faas BH, **de Ligt J**, Janssen I, Eggink AJ, Wijnberger LD, van Vugt JM, Vissers L, Geurts van Kessel A. Non-invasive prenatal diagnosis of fetal aneuploidies using massively parallel sequencing-by-ligation and evidence that cell-free fetal DNA in the maternal plasma originates from cytotrophoblastic cells. Expert Opin Biol Ther. 2012; 12 Suppl 1:S19-26.

Neveling K, Collin RW, Gilissen C, van Huet RA, Visser L, Kwint MP, Gijsen SJ, Zonneveld MN, Wieskamp N, **de Ligt J**, Siemiatkowska AM, Hoefsloot LH, Buckley MF, Kellner U, Branham KE, den Hollander AI, Hoischen A, Hoyng C, Klevering BJ, van den Born LI, Veltman JA, Cremers FP, Scheffer H. Next-generation genetic testing for retinitis pigmentosa. Hum Mutat. 2012; 33(6):963-72.

Itsara A, Vissers LE, Steinberg KM, Meyer KJ, Zody MC, Koolen DA, **de Ligt J**, Cuppen E, Baker C, Lee C, Graves TA, Wilson RK, Jenkins RB, Veltman JA, Eichler EE. Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. Am J Hum Genet. 2012; 90(4):599-613.

Willemsen MH, Vissers LE, Willemsen MA, van Bon BW, Kroes T, **de Ligt J**, de Vries BB, Schoots J, Lugtenberg D, Hamel BC, van Bokhoven H, Brunner HG, Veltman JA, Kleefstra T. Mutations in *DYNC1H1* cause severe intellectual disability with neuronal migration defects. J Med Genet. 2012; 49(3):179-83.

Vissers LE*, **de Ligt J***, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA. A *de novo* paradigm for mental retardation. Nat Genet. 2010; 42(12):1109-12.

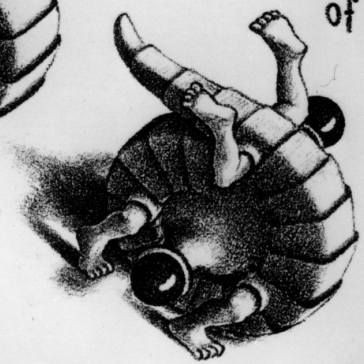
* shared first authorship

Curriculum Vitae

Joep de Ligt was born on the 20th of September 1986 in Boskoop, the Netherlands. After completing the HAVO in 2003 he started his applied bachelor education in bioinformatics at the Hogeschool Arnhem Nijmegen in Nijmegen. During this period he did internships at the Centre for Molecular and Biomolecular Informatics and N.V. Organon on protein flexibility, stability and ligand interactions. Upon finishing his Bachelor of Science in 2007 he was admitted to the Master of Molecular Life Sciences program at the Radboud University Nijmegen. Several courses in this Master were selected from the curriculum of the Bioinformatics Master program at the Wageningen University.

Upon completion of his Master thesis on rotamers in 2009 he started as a PhD student under supervision of Joris Veltman and Han Brunner at the Radboud University Medical Center in Nijmegen at the department of Genetics. Here he performed research on the genetics underlying the sporadic occurrence of intellectual disability. He received several awards for this work amongst which are: a semi-finalist placement for the young investigator awards at the American Society of Human Genetics (ASHG) in 2012 and the European Society of Human Genetics (ESHG) Venniea Medical Academy award in 2013. During this time he was also able to attend the Leena Peltonen School of human genomics in 2012 and visit the prestigious lab of James Lupski at the Baylor College of Medicine in Houston, Texas, USA, for three months for a collaborative project on rare genomic disorders in 2013. He finished his work for his thesis in November 2013 upon which he started as postdoctoral researchers at the Hubrecht institute in Utrecht within the group of Edwin Cuppen on the topic of genomic stability and cancer.

De Pedalternorotandomovens centroculatus articulosus ontstond, (generatio spontanea!) uit onbevredigdheid over het in de natuur ontbreken van wielvormige, levende schepselen met het vermogen zich rollend voort te bewegen. Het hierbij afgebeelde diertje, in de volksmond genaamd „wentelteefje" of „rolpens", tracht dus in een diepgevoelde behoefte te voorzien. Biologische bijzonderheden zijn nog schaars: is het een zoogdier, een reptiel, of een insekt? Het heeft een langgerekt, uit verhoornde geledingen gevormd lichaam en drie paren poten, waarvan de uiteinden gelijkenis vertonen met de menselijke voet. In het midden van de dikke, ronde kop, die voorzien is van een sterk gebogen papagaaiensnavel, bevinden zich de bolvormige ogen, die, op stelen geplaatst, ter weerszijden van de kop ver uitsteken. In gestrekte positie kan het dier zich, traag en bedachtzaam, door middel van zijn zes poten, voort bewegen over een willekeurig substraat (het kan eventueel steile trappen opklimmen of afdalen, door struikgewas heendringen of over rotsblokken klauteren). Zodra het echter een lange weg moet afleggen en daartoe een betrekkelijk vlakke baan tot zijn beschikking heeft, drukt het zijn kop op de grond en rolt zich bliksemsnel op, waarbij het zich afduwt met zijn poten, voor zoveel deze dan nog de grond raken. In opgerolde toestand vertoont het de gedaante van een discus-schijf, waarvan de centrale as gevormd wordt door de ogen-op-stelen. Door zich beurtelings af te zetten met één van zijn drie paren poten, kan het een grote snelheid bereiken. Ook trekt het naar believen tijdens het rollen (b.v. bij het afdalen van een helling, of om zijn vaart uit te lopen) de poten in en gaat „freewheelende" verder. Wanneer het er aanleiding toe heeft, kan het op twee wijzen weer in wandel-positie overgaan: ten eerste abrupt, door zijn lichaam plotseling te strekken, maar dan ligt het op zijn rug, met zijn poten in de lucht en ten tweede door geleidelijke snelheidsvermindering (remming met de poten) en langzame achterwaartse ontrolling in stilstaande toestand.



XI-'51

Mil Luther Proefdruk

Dankwoord

DANKWOORD

Van harte welkom bij het dankwoord, je hebt het gehaald, het feit dat je dit leest betekent dat je in meer of mindere mate verantwoordelijk bent voor het resultaat dat voor je ligt. Of het nu is als een begeleider, als collega die het veld gebracht heeft naar waar het nu is, als liefhebbende familie of partner of als analist die het 'echte' werk gedaan heeft, allemaal was het nodig, heel erg bedankt hiervoor. Nu zijn er natuurlijk wel een aantal mensen die ik hier graag even extra in het zonnetje zet.

Genetica Nijmegen is een top afdeling, niet alleen qua wetenschap maar ook qua mensen, ik ben oprecht vereerd dat ik hier mijn PhD heb kunnen doen, iedereen die de afdeling maakt tot wat die is, bedankt.

Han, dank voor het vormgeven van een geweldige afdeling en het mogelijk maken van topsport in de wetenschap. Verder hebben jouw vertrouwen, oprechtheid en altijd prikkelende vragen mijn PhD tot een succes gemaakt.

Joris, jij zag het wel zitten om een bioinformaticus op een 'lab' project aan te nemen, genetica werd tenslotte toch een computer vak. Daar ben ik nog steeds blij mee en ik denk dat we iedereen hebben laten zien hoe een goede combinatie tussen 'wet' en 'dry' lab tot prachtige wetenschap leidt. Bedankt voor je altijd scherpe commentaar en dat je me zo af en toe weer even met beide benen op de grond zette.

Lisenka, zonder jouw was er geen data geweest, ik benijd nog steeds jouw onfeilbare organisatie. Bedankt voor al je input en de prachtige data die je hebt weten te generen. Die enkele sample swaps die we hadden, daar kwamen we wel uit, hoe ingewikkeld ze ook waren.

Jayne, wat Lisenka op het lab voor elkaar krijgt dat lukt jou bij de bioinformatica, van een zootje ongeregeld maak jij een georganiseerde groep mensen, en dat is niet niks. Het leukste waren nog de puzzeltjes van Rolph of later uit je eigen data, zeker op 'sanity optional' vrijdagen.

Dan hebben we nog de twee mooi-boys, mijn vrienden en voor een dag 'mijn' paranimfen **Lex** & **Eugène**. Jullie wil ik boven alles bedanken voor de gezellige tijden, goede gesprekken en de lol. Maar natuurlijk ook voor het nu al onvergetelijke dansje.....

Tot dusver de overduidelijke kandidaten, maar onderzoek als dit doe je niet alleen, zeker dit onderzoek niet. **Alex** en **Christian**, jullie hebben het grondwerk gedaan

waarop ik voort kon bouwen, zonder jullie was het nooit zo soepel gelopen. Dan natuurlijk de mensen die het minder vernieuwende maar minstens zo noodzakelijke werk gedaan hebben. **Nienke**, **Marisol**, **Eugène** & **Rick**, sorry voor alle 'rotklusjes' maar dankzij jullie had ik veel meer tijd om wetenschap te doen, bedankt daarvoor. **Djie**, ik ben er nog steeds heel blij om jouw begeleid te hebben, wat een mooi resultaat, en straks ook nog een paper.

Dan hebben we natuurlijk ook nog ons meer dan fantastische lab team, **Petra**, **Marloes**, **Irene**, **Thessa**, **Michael**, **Peer** & **Bart**. Thessa jou heb ik meer dan een beetje stress bezorgt, maar gelukkig niet voor niets. Jullie allen wil bedanken voor je inzet en de gezelligheid die jullie met de lunch, uitjes en feestjes altijd mee brachten. Hier droeg de 'onco' groep (**Simon**, **Richarda**, **Robbert**, **Ingrid**, **Esme**, **Eveline** & **Marc**) ook zeker aan bij, het was gezellig. **Rocío**, **Bonnie**, **Susanne**, **Merel**, **Tom**, **Arjen** & **Cees**, bedankt voor jullie vragen en input tijdens besprekingen en presentaties.

Een zeer belangrijke dataset was de microarray dataset, **Rolph** & **Nicole**, bedankt voor jullie uitleg en hulp en de interessante cases en de **array diagnostiek** voor het draaien van de arrays. Al het onderzoek in dit proefschrift was niet mogelijk geweest zonder de uitgebreide en zorgvuldig geselecteerde patiënten verzameling, hiervoor mijn grote dank aan de clinici en diagnostiek, **Marjolein**, **Bregje**, **Anneke**, **Bert**, **Tjitske**, **Janneke**, **Helger** & **David**. Onderzoek gaat ook niet zonder funding, **Hans** & iedereen bij TechGene bedankt.

Dear **Jim** and **Lupski lab members**, thank you for the great science and making me feel at home those 3 months I spent in Houston, I whish you all the best. **Jim**, thank you for having me over and the experience of being part of a top US research lab. **Claudia**, **Christine** & **Ian**, you showed me some truly awesome hospitality and I hope to be able to repay the favor some day. To **'the Ginger Man' crowd**, thank you for the great beers and even greater company, keep colleting those glasses. **Richard**, tank you for your input and your slightly unsettling but fun insight into Aussie culture.

Nog even terug komend op het thema gezelligheid, hier ga ik niet iedereen bij naam noemen maar je weet dat ik jullie bedoel. Jullie die bij **Aesculaaf** nog een drankje deden of kwamen **borrelen** voor een paper, jullie die mee gingen **poolen**, **pokeren** en **films kijken**, jullie die bij de NCMLs **PhD retreat** mij vertrouwde om een kroeg te vinden, jullie die er bij de **Batavieren race**, de **7 heuvelen loop** en de **triatlon** waren, jullie die op **congres** mee uit gingen, allemaal bedankt! Zonder gezelligheid geen wetenschap voor mij.

Filosofische gedachte vloeide rijkelijk tijdens onze bijeenkomsten en anders had Iemand wel een slechte grap. De drive om te winnen was groot en het spel was Fantastisch, behalve als het van Persie was, maar daar kon hij ook niks aan doen. Amateurs dat waren we ooit, nu zijn we volleerd, dames (**REMBD**) deze is voor jullie, de volgende keer weer bij mij?

Tim, **Tom**, **Tom2** en **Lex**, Luxemburg en Bilbao brengen nog altijd een glimlach op mijn gezicht. Ik ben benieuwd wat de bestemming dit jaar wordt, het wordt weer eens tijd.

En dan natuurlijk mijn **familie**, alle ooms, tantes, neven en nichten die ieder familie weekend weer geïnteresseerd waren in wat ik deed en geduldig naar mijn soms erg omslachtige uitleg luisterde. Bedankt voor jullie vragen en enthousiasme.

Lieve **grootouders** jullie wil ik bedanken voor een fantastische jeugd en het opvoeden van mijn lieve ouders, **Opa**, jou dank ik in het bijzonder, de reden dat ik dit boek aan jou opdraag is omdat jij van iedereen vaak de beste vragen stelde, niet alleen over mijn onderzoek maar ook over wetenschap in het algemeen, iets wat mij heeft geleerd altijd vragen te durven stellen.

Lieve **Marjoke**, **Paul** & **Ninge**, afgezien van het feit dat ook jullie altijd geïnteresseerd en lief zijn hebben jullie ervoor gezorgd dat Jeuske is wie ze is, en daar ben ik jullie eeuwig dankbaar voor.

Lief schattig zusje (sorry kon het niet laten), lieve **Marlies**, altijd druk met studeren, en niet zonder resultaat, ook jij mag straks zo'n leuk 'boekje' gaan schrijven, maar dat komt vast goed. Bedankt voor al die gezellige vakanties en af en toe een bed om in te slapen.

Lieve **Els** en **Sjaak**, of jullie het nu willen of niet, jullie zijn grotendeels verantwoordelijk voor dit boek, door jullie ben ik wie ik ben. Dank voor jullie steun in al die jaren, ook toen het wat minder ging bleven jullie er in geloven dat ik het kon, bij deze bewijs ik graag dat jullie gelijk hadden.

Einstein & **Newton**, jullie zijn altijd een inspiratie en warmte bron voor mij geweest, bedankt voor alle kopjes en gespreksstof.

Dan de laatste, mijn alles, mijn schat, **Jeuske**, zonder jou was ik nooit zo gelukkig geweest als dat ik ben, en dat is behoorlijk gelukkig. Dank je voor je begrip als ik weer eens laat was of naar het buitenland moest, bedankt voor je goede zorgen terwijl ik ziek was of het weer eens heel druk had. Bedankt voor het grootste geluk in de wereld, kleine **Lerris**, ons gezinnetje, heerlijk.